SJIF Impact Factor 2022: 8.197| ISI I.F. Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 7 | Issue: 4 | April 2022 - Peer Reviewed Journal

ONLINE FRAUD DETECTION

Sonia Chhabra[1], Vaibhav Gupta[2], Amit Kumar[3], Aman Anand[4]

Department of Computer Science and Engineering, Sharda University
Greater Noida, Uttar Pradesh - 201310

Article DOI: https://doi.org/10.36713/epra9979
DOI No: 10.36713/epra9979

ABSTRACT

In today's fast moving world as the demand of the internet is increasing with this cyber-attacks are also increasing and Phishing is one the most common cyberattack among them. Taking a user's personal information such as credit card numbers, login details, confidential info, and so on through illegitimate activities of using the internet without the user's knowledge and using it for blackmailing, debiting money from the user's account, or any other purpose with the wrong intentions is known as phishing. In phishing attack, a phisher or attacker pretends to be masquerade as a known person or organization to the user and sends the mails or messages which contains malicious links in them and these malicious links contains harmful software's or viruses which steals the user's computer data, financial data, login credentials such as User ID and passwords, credit card details, etc. Phishing is the most common and dangerous cyberattack which is growing in the today's world. Nowadays phishers are working smartly they are using the new techniques for creating the malicious links and embeds them in the emails and messages and sends it to the user which looks similar to the trusted mail or message to the user and as soon as the user clicks on the malicious link it redirects the user to the malicious webpage or runs the harmful software in the backend while the user is reading the email or message and takes over the control of the user's computer and steals all data of the user's computer. Due to a speedy development inside the digital commerce generation, the usage of credit playing cards has dramatically extended. In view that credit card is the most popular mode of fee, the number of fraud instances related to it is also rising. As a result, in order to prevent these frauds, we need an excellent fraud detection system that can detect them correctly. We created the idea of credit card frauds in this paper, and we used a variety of device learning methods on an unbalanced dataset, including logistic regression, naivebayes, and random wooden area with ensemble classifiers using the boosting approach. An in-depth analysis of the existing and proposed models for credit card fraud detection has been completed, as well as a comparison of these tactics. So one of a kind classification models are applied to the statistics and the model performance is evaluated on the basis of quantitative measurements which include accuracy, precision, recollect, f1 score, confusion matrix. The realization of out observe explains the first class classifier via schooling and trying out using supervised strategies that offers better answer.

KEYWORDS: *phishing, credit card, cyberattack, malicious, detection, webpage.*

I. INTRODUCTION

Annually, phishing websites cost web users, businesses, and organisations tens of billions. because they steal their personal data, financial data, login credentials such as User Id's and passwords, as a result these companies and organizations are going into loss which affects not only the company but also it affects more badly to their employees because they lose their jobs when any company goes into loss. In the current Covid-19 pandemic the use of the internet has increased very rapidly as a result all the physical activities such as official meetings, classes, shopping, financial activities, etc. shifted from physical mode to online mode which gives the opportunities to phishers to attack on the user's data. Companies are also promoting work from home due to the guidelines of the government, schools and colleges has shifted offline classes to online mode to avoid mass gathering and for safety reasons, in fact every sector is working in online mode which increases the cyberattacks and most common and dangerous one is the phishing attacks. In phishing attack phishers or attackers sends emails and messages to the user's and pretends as if they are from a genuine organization or a person and as soon as the victim clicks on that email or message which contains malicious links it starts their process in their backend or redirects the user to the some other page which contains some type of forms which asks for the user's identity and as soon as user's fills all the details including credit card details all the information of the user goes to the attacker who uses this information for blackmailing purpose and flushes all the money from the user's account and the victim loses all his/her hard earned money in just a minute. Nowadays it is becoming difficult even for the cybercrime department to track the phisher or keep track on phishers because they are also smart enough because they are also engineer's and well educated persons who are having a good knowledge of computer's and new technologies and they create those kind of malicious links which looks similar to a genuine website and hard to detect whether it's a genuine or a fake one. As a result for the detection of these type of links it becomes more important to use new kind of approach which can detect these links which can't be detected by the older approach. Detection of phishing websites is not an easy task because URLs are altered in a variety of ways, including shortened URLs, link redirections, and modifying links to make them appear trustworthy, to name a few. This necessitated a shift away from traditional programming methods and toward machine learning. In recent years, humans have been concerned about the winning statistics mining model, which is mostly based on statistics mining. Classification data mining algorithms aren't immediately appropriate because our problem is handled as a type problem. Supervised learning algorithms are evolutionary algorithms that aim to provide better results as time goes on. Credit card is the most popular mode of price because the number of credit card users is rising very hugely, the identity robbery is multiplied, and frauds also are increasing, within the virtual card purchase, best the card record is needed which includes card variety, expiration date, comfortable code, and so forth. Such purchases are commonly achieved on the net or over phone. To commit fraud in those forms of purchases, a individual surely desires to know the card information, The mode of charge for online buy is in general achieved by using credit score card. The information of credit card need to be saved personal. To secure credit score card privateness, the details ought to now to be leaked. Exceptional ways to scouse borrow credit score card information are phishing sites, steal/lost credit score playing cards, counterfeit credit score cards, robbery of card info, intercepted cards and many others. For security reasons, the above matters need to be averted. The purchase is made remote best the card's data are wished in online fraud. At the time of purchase, a guided signature, a PIN, or a card imprint are not required. In maximum of the cases the real card holder is not conscious that someone else has seen or stolen his/her card records. The simple manner to locate this sort of fraud is to analyze the spending styles on every card and to figure out any variant to the traditional spending patterns. Fraud detection by analyzing the existing facts purchase of cardholder is the first class manner to decrease the rate of a success credit score frauds. Because these facts sets aren't available and additionally the outcomes are not disclosed to the public. The fraud cases should be the detected from the available record sets referred to as the logged records and person conduct. At present, the online fraud detection has been implemented by using some of techniques including information mining, records and synthetic intelligence.

II. PHISHING DETECTION USING URL

As we are all aware about the vast usage of internet which involves browser surfing (visiting websites). This is one of the major reasons where phishers target to steal personal information of user. Different types of algorithms are employed to prevent various kind of attacks (ML algorithms).

A. PHISHING

Phishing is a form of cyber bullying in which phishers obtain the user's credentials without their knowledge. It is a combination of technology and social engineering that is used to collect personal and private information such as credit card numbers and other financial information. There are various ways of cyber attacks using phishing , some of them are listed below:

1. An illegal website is created which resembles just like original website.
2. Phishers sends various text messages in name of verified sender along with that phishing website link, and user unknowingly visits their website.
3. The user accesses the website and provides his or her personal information by clicking on the link provided by the sender.
4. Phishers then use that information in carrying out other illegal stuffs.

B. APPROACH TO MACHINE LEARNING

In this approach, URL features are obtained and used to determine whether a website is phishing or not.In this various ML algorithms are used which work on features extracted from URL.

1. Naïve Bayes algorithm

The Nave Bayes algorithm is a supervised machine learning method used mostly for classification tasks. A machine learning model known as Nave Bayes is utilised to make fast predictions. It predicts the result based on the likelihood of the thing. Because it uses the Bayes' Theorem principle, it's called Bayes. The Nave Bayes method is a machine learning approach for forecasting a set of datasets that is straightforward to implement.

It can be used to classify binary and multi-class data. It's a popular tool for text classification. Credit rating, spam detection, and medical data classification are just a few of the applications.

2. K-Nearest Neighbor Algorithm

K-Nearest Neighbors is a type of supervised learning method that is based on the simplest machine learning algorithms. The K-Nearest technique is utilised in a variety of applications, including data mining, pattern recognition, and intrusion detection. The K-nearest approach can be used to tackle both classification and regression issues. K-Nearest Neighbor is built on the measure of similarity, and it stores all available examples before generating a new set of cases based on the

similarity measure.
Because it uses all available data for training purposes during classification, K-Nearest Neighbors is also known as the Lazy Learning Algorithm. KNN is utilized in regions where there is significant domain knowledge, which is why it is employed in applications that require high accuracy. When a big collection of training data is available, KNN can be employed efficiently.

3. Decision Tree Algorithm

The category of Supervised machine learning algorithms includes decision trees. It can be used to tackle problems involving classification and regression. In a decision tree, there are two sorts of nodes: Decision Node and Leaf Node. Decisions have several branches and are used to make decisions, whereas Leaf Nodes have no branches and indicate the decisions' result. The structure of a Decision tree is similar to that of a tree, which is why it is called a Decision tree. It features a root node that has many branches and resembles a tree.

The decision tree is simple to comprehend because it makes decisions in the same way that humans do. If-else conditions are comparable to decision trees. It tests the condition first, and if it's true, it moves on to the next node, which makes the prediction.

4. Random Forest Algorithm

The Random Forest method is based on the technique of Supervised Machine Learning. It's utilised for both classification and regression problems. Random forest creates decision trees from a variety of datasets and then votes on the best answer from those trees. Random forest chooses a variety of random samples from the dataset. Then it constructs a decision tree for each sample and obtains the predicted result for each decision tree, votes on those predicted results, and finally chooses the most voted forecast result as the final result. In this approach, it is able to forecast the outcomes with more precision. It solves the problem of overfitting by averaging the different decision trees.

Random forest is capable of handling extremely huge datasets. In circumstances where a major amount of the data is absent, random forest produces accurate results.

5. Support Vector Machine Algorithm

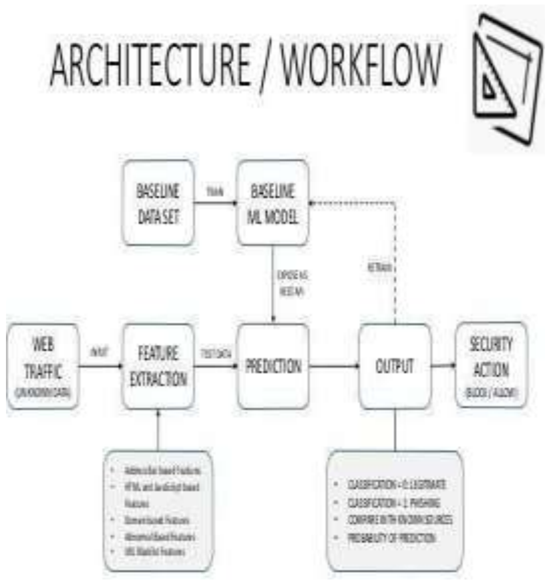
The Support Vector Machine is based on the Supervised Machine Learning algorithm, which is used for both classification and regression of issues. SVM creates a decision boundary that divides the n-dimensional space into classes, making it easier to categorise new data points later. A Hyperplane is the name for this decision boundary. SVM selects extreme points to aid in the construction of the hyperplane. Support vectors represent the extreme examples, which is why the technique is called Support Vector Machine.

Because it uses a subset of training points in the decision function known as support vectors, SVM is a memory-efficient approach. SVM provides high-accuracy findings and works well in high-dimensional spaces.

III.PROPOSED WORK

1. Developing a browser plugin that can track all of a user's system traffic (http). We developed a browser extension rather than software or an application to ensure that the system is entirely real-time. [1]
2. The next stage is a comparison-based task in which the URL is compared to URLs on the whitelist and blacklist. Web scraping will be utilised to extract the data for these lists on the fly. If the URL's domain falls into the whitelist category, mark it as safe; otherwise, proceed to the next step and use other methods. [1]

3. The website will now be thoroughly examined utilising the various functionalities. We used the following features: URL length, favicon similarity, website registration and expiration dates, number of @ symbols used in the URL, number of dots used in the URL, website protocol (secure or unsecured), number of hyphens (-) used in the URL, and whether the URL uses direct IP address or not. [1]
4 If the hyphen in the URL equals 1, the website is suspect. If the URL < 1 has a hyphen, it is a legitimate website. Because attackers increasingly design malicious webpage that seem like trusted web pages, we extract and compare the CSS of the suspicious URL with the CSS of the real URL in the next technique. [1]
5. Then we apply random forest, decision tree, and logic regression machine learning algorithms to the collected data to generate the score. [1]
6. Then we calculate the similarity and match score, and if it is more than the threshold, the URL is marked as phishing and blocked.
7. The aforementioned technique is more efficient, precise, and safe than any other initial solution since it creates a three-level security barrier.


ARCHITECTURE / WORKFLOW
BASELINE DATA SET
TRAIN
BASELINE ML MODEL
EXPOSE AS REST API
RETRAIN
WEB TRAFFIC (UNKNOWN DATA)
INPUT
FEATURE EXTRACTION
TEST DATA
PREDICTION
OUTPUT
SECURITY ACTION (BLOCK / ALLOW)
Address Bar based Features
HTML and JavaScript based Features
Domain based Features
Abnormal Based Features
URL Blacklist Features
CLASSIFICATION = 0: LEGITIMATE
CLASSIFICATION = 1: PHISHING
COMPARE WITH KNOWN SOURCES
PROBABILITY OF PREDICTION

IV.ANALYSIS PHASE

The following are the detailed rules that we devised as a result of our research:

1. Domain length in URL

If the length is between 3 and 20, it is considered genuine.

If the length is between 20 and 24, be cautious. If the length exceeds 24, it has been phished. [1]

2. The @ sign in the domain

If the amount of @ symbols is zero, the website is real; otherwise, it has been phished.. [1]

3. The keyword "http" is used between domains.

If "http" appears in the domain, it is phished; else, it is legal. [1]

4. Is there a protocol?

If yes, it is legitimate; otherwise, it is suspect. [1]

5. If the time gap between the expiration date and the date of registration on the website is higher than 90 days, the website is legitimate; otherwise, it is suspect. [1]

6. IP Address (Direct)

If the URL contains a numeric value in IP address, it is suspect; otherwise, it is valid.

7. In a domain, the number of hyphens is

If the number of hyphens is zero, the website is real; otherwise, it has been phished.

8. Google indexing for favicon similarity

The website has been phished if the favicons of the two websites are identical but the domain of the URL is distinct.

CONCLUSION

The suggested system, which is built using the above approaches, is safer and more effective than prior systems since it allows the user to safely surf the web and all transactions conducted by the user using the above system are secure. The above system ensures that the user's private information is kept private. It is much easier to provide our proposed system to users in the form of browser extensions. Because we applied machine learning algorithms, the proposed solution is extremely efficient. The attackers are continually devising new techniques to counter our proposed solution, which is one special issue.

To solve this challenge, we need new algorithms that can adapt to the characteristics of phishing URLs. The system that will be created with the above algorithms will be more precise. The system will be more accurate, efficient, and secure and protected if all of these diverse ways are combined. The weaknesses in the system can be solved by giving a much richer characteristic to the machine learning model, which would result in much greater accuracy.

REFERENCES

1. *Vaibhav Patil, Pritesh Thakkar, Cjirag Shah, Tushar Bhat, Prof. S.P.Godse, "Detection and Prevention of Phishing Websites using Machine Learning Approach", 3rd ed., vol. 2.*
2. *Oxford: Claren*
3. *Mahajan Mayuri Vilas, Kakade Prachi Ghansham, Sawant Purva Jaypralash, "Detection of Phishing Website Using Machine Learning Approach" 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT).*
4. *M. Srinithya , V. Ragasree , K. Sri Harshini ,"Phishing Websites Detection" International Journal of Science and Research (IJSR).*
5. *Bandeh Ali Talpur, Declan O"Sullivan ,"Cyberbullying severity detection: A machine learning approach",*

research article PLOS ONE.

6. *Srushti Patil, and Sudhir Dhage, "A Methodical Overview On Phishing Detection Along With An Organized Way To Construct an Anti- Phishing Framework", 2019 5th*
7. *International Conference On Advanced Computing ·Communication System(ICACCS), pp. 1-6.*
8. *Huaping Yuan, Xu Chen, Yukun Li, Zhenguo Yang and Wenyin Liu, "Detecting Phishing Websites and Targets Based On URLs and Webpage Links", 2018 24th International Conference on Pattern Recognition(ICPR) Beijing, China, August 20-24, 2018.*