International Journal of Interdisciplinary Innovative Research &Development (IJIIRD)
ISSN: 2456-236X
Vol. 05 Issue 02 | 2021

Recognizing Phishing Sites Using ML

[1]Mr. Maheshwar M, [2]Dr Gangothri Rajaram
[1]*MCA Scholar, School of CS & IT, Dept of MCA, Jain (Deemed-to-be) University, Bangalore*
[2]*Asst. Professor, School of CS & IT, Dept of MCA, Jain (Deemed-to-be) University, Bangalore*

ABSTRACT

Phishing is the strategy of extracting user credentials and sensitive data from users by masquerading as a genuine website. In phishing, the client is given a mirror site that is identical to the authentic one yet with malicious code to remove and send client certifications to phishers. Many users unwittingly click phishing domains every day and every hour. The attackers are targeting both the users and the companies. For ex. Microsoft remains the #1 spoofed brand in phishing attacks due to the growing adoption of Microsoft 365. Between Q1 and Q3 2020, Vade Secure detected 30,608 unique Microsoft phishing URLs. The objective of my project is to execute an ML solution to detect phishing and malicious websites. Here we are going to implement regression Algorithms to detect the Phishing Websites & Legitimate Websites & going to compare the build model with different types of regression Algorithms like Logistic Regression, Random Forest & (SVM) Support Vector Machine Algorithm. We are performing this operation from the dataset gathered from the Kaggle.

Keywords: ML, Phishing, Logistic Regression, Random Forest Classifier, Microsoft, Legitimate, Websites, Support Vector Machine Algorithm.

1. INTRODUCTION

Phishing is a form of fraud in which the attacker tries to learn sensitive information such as login credentials or account information by sending as a reputable entity or person in email or other communication channels. Phishing is famous among attackers, since it is simpler to fool somebody into clicking a malicious link which appears to be genuine than attempting to get through a PC's safeguard frameworks. The malicious link inside the body of the message is intended to cause it to give the idea that they go to the spoofed association utilizing that association's logos and other genuine contents. Features used to detect Phishing Websites: Uniform Resource Locator (URL): URL is the primary thing to dissect a site to choose whether it is a phishing or not. URLs of phishing areas have some particular focuses. Features which are identified with these focuses are gotten when the URL is processed. Typo squatting: is a type of cybersquatting that depends on mistakes, for example, typographical errors made by Internet users while entering the site address into an internet browser or dependent on typographical mistakes that are difficult to see while fast perusing. URLs that are made with Typo squatting resemble a trusted domain. A client may coincidentally enter an incorrect website address or click a link that looks like a trusted domain, and in this way, they may visit an alternative website owned by a phisher. Cybersquatting: is registering, trafficking in, or using a domain name with bad faith intent to profit from the goodwill of a trademark belonging to someone else. The cyber-squatter may offer to sell the domain to a person or company who owns a trademark contained within the name at an inflated price or may use it for fraudulent purposes such as phishing. This can be reduced with the help of Machine Learning Models. Machine Learning is a strategy for data examination that automates analytical model structure. It is a part of AI dependent on the possibility that frameworks can gain from information, distinguish patterns and settle on choices with minimal human interaction. Logistic Regression is utilized to make the predictions of event is Success (1) or Failure (0). It's implemented when the Dependent Variables are in Binary Format (0/1).

Random Forest is a technique that works by developing different decision trees. A ultimate conclusion is made dependent on most of the trees and is picked by the Random Forest. A Decision tree is a tree-molded diagram used to decide a strategy. Each part of the tree addresses a potential choice, event, or response.

A SVM stands for Support Vector Machine model is fundamentally a representation of various classes in a hyperplane in multidimensional space. The hyperplane will be created in an iterative way by Support Vector Machine (SVM) Algorithm with the goal that the blunder can be limited. The objective of Support Vector Machine (SVM) Algorithm is to separate the datasets into classes to track down a Maximum negligible hyperplane.

Here we are going to use this Logistic Regression, Random Forest & Support Vector Machine (SVM) Algorithm to detect the phishing Website & make the comparison between these three & find the best algorithm. We Have collected the Dataset from the Kaggle website with 31 features and we are going to use python and develop a fully working website which detects the phishing website.

2. LITERATURE REVIEW

[1] The Authors focus is to pursue a higher performance classifier by studying the features of phishing website and choose the better combination of them to train the classifier. We study all features to indicate the strongest, weakest and to remove the irrelevant features; the study is based on examining all possible combination of the available features. The system acts as an additional functionality to an internet browser as an extension that automatically notifies the user when it detects a phishing website. The system is based on a machine learning method, particularly supervised learning. We have selected the Random Forest technique due to its good performance in classification. The experimental results of using Fuzzy logic algorithms were unexpectable here some experiments had approximately 100 % of accuracy rates by applying only five features.

[2] The author focuses on detecting phishing website URLs with domain name features. Web spoofing attack categories content-based, heuristic-based and blacklist-based approaches. As technology continues to grow, phishing techniques started to progress rapidly and this needs to be prevented by using anti-phishing mechanisms to detect phishing. Machine learning is a powerful tool used to strive against phishing attacks. The author has made a Survey on 10 papers listed the some of the best techniques like Random Forest, ELM (Extreme Learning Machine), Neural Networks & etc.

[3] In This paper the author has made a study and by gathering the exploration papers and said centers around distinguishing phishing site URLs with domain names. Spoofing attack classifications content-based, heuristic-based and boycott-based methodologies are clarified and the proposed model Phish Checker is created with the assistance of Microsoft Visual Studio Express 2013 and C# language. The had made a study on the URL features in order to reduce time computation and providing high performance with the least combination of the powerful features. However, because of time shortage and hardware limitation, we chose random features to process its combination. We concluded after some observation that the combination of features computed take the shape of normal distribution curve, it starts with least combination of features with low probability of combination and time consuming, then picks up accordingly, then goes down as it reaches final number of features.

[4] In order to assist the user to be aware of the access to such websites, the implemented system notifies the user through email. Author in this Paper explained exceptional highlights, for example, catching blacklisted URLs from the program straightforwardly to confirm the legitimacy of the site, advising client on blacklisted sites while they are attempting to access through pop-up, and furthermore informing through email. This framework will help client to be ready when they are attempting to get to a blacklisted site. The Author has said the if we find any phishing sites it would be good to block the site and alert the use through the email so that the user will have the track of the activities.

[5] The classifiers were utilized to recognize phishing URLs. In identifying phishing URLs, there are two stages. The initial step is to separate highlights from the URLs, and the subsequent advance is to arrange URLs utilizing the model that has been created with the assistance of the preparation set information. In this work, we utilized the data-set that gave the extracted features. The data-set, from The University of California, Irvine Machine Learning Repository, contained nine features. The research work presented here has some limitations and it can be extended further. The first limitation is that we considered a small data set that contains 1353 URLs, and there are 9 features for each URL. The second limitation is that all features are discrete.

[6] The possibility that author is advancing here is to improve the productivity by utilizing Random forest as our classification algorithm with the assistance of R studio tools that helps us in better examination. Parsing is done to analyze feature set. We confine our features to 8 out of the 31 features that are considered by parsing and by thorough rigorous analysis. Here, parsing is done using Attribute Subset Selector which incorporates two sections 1) Attribute Subset Evaluator Algorithm utilizing Information Gain 2) Search Method Algorithm utilizing Ranker Method. Parsing is executed using a java code which imports WEKA tools libraries for IG and Attribute selector. Those Attribute that offers more data will have a higher data acquire esteem and can be chosen, though those that don't add a lot of data will have a lower score and can be taken out. In this paper, a different methodology has been proposed to detect phishing websites by using random forests the performance metrics along with literature survey also proved the accuracy level of random forest to be the around 95% and thus Random Forests were chosen for classification.

[7] This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. This paper plans to improve identification strategy to distinguish phishing sites utilizing AI innovation. Hear the creator has accomplished high exactness utilizing arbitrary woodland calculation with least bogus positive rate. Likewise result shows that classifiers give better execution when we utilized more information as preparing information. Based on the research and review the author maid the concludes that Random Forest has the Best outcome in finding the phishing URLs.

[8] To successfully identify phishing assaults, this paper planned another identification framework for phishing sites using LSTM Recurrent Neural Networks (RNN). LSTM has the upside of catching data timing and

long-term dependencies. LSTM has solid learning capacity, can naturally learn data characterization without manual extraction of complex features, and has solid potential despite complex high-dimensional massive data. Trial results show that this model methodology the precision of 99.1%, is higher than that of other neural organization calculations.

[9] The point is to carry out the location of the phishing sites using data mining. This assignment is done by extracting the features of the site through URL when the client visits it. The obtained features will go about as test information for the model. Random Forest Algorithm can be utilized to prepare the proposed model. The principle undertaking of this framework is to recognize the phishing site and caution the client heretofore in order to keep the clients from getting their qualifications abused. On the off chance that any client actually wishes to continue, it tends to be done at their own danger. The main task of this system is to detect the phishing website and alert the user beforehand so as to prevent the users from getting their credentials misused. If any user still wishes to proceed, it can be done at their own risk.

[10] In this paper to evaluate the performance of phishing URL detection, the Author has randomly chosen 100 phishing URLs from Phish Tank for evaluations. The proposed phishing objective discovery approach has discovered right phishing focuses for 78 phishing URLs, wrong focuses for 5 phishing URLs, while the 17 excess ones can't be gotten too. All the more explicitly, the proposed search administrator-based technique for phishing objective identification accomplishes a precision of 93.98%. The exploratory outcomes demonstrate the adequacy of the pursuit administrator based phishing objective recognition approach.

3. PROBLEM STATEMENT

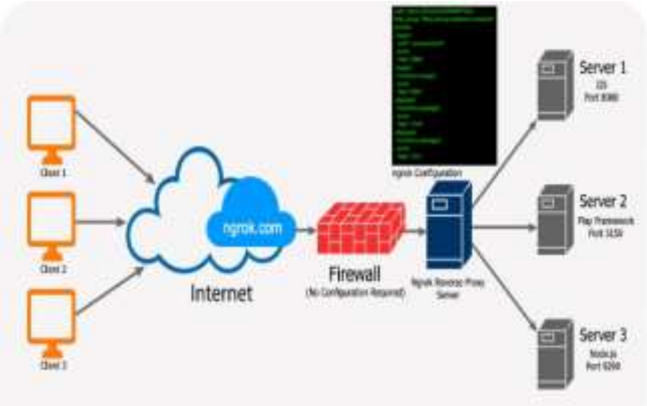
Phishing imitates the characteristics and features of emails and makes it look the same as the original one. It appears similar to that of the legitimate source. The user thinks that this email has come from a genuine company or an organization. This makes the user to forcefully visit the phishing website through the links given in the phishing email. The phishers force user to fill up the personal information by giving alarming messages or validate account messages etc. so that they fill up the required information which can be used by them to misuse it. They make the situation such that the user is not left with any other option but to visit their spoofed website.

4. SOLUTION

The proposed solution is used to detect the phishing web sites based on many features which are grouped as: Domain-Based Features, URL-Based Features. Logistic Regression is used to predict the event success or not. When dependent variables are in binary form (0/1) Logistic Regression is used.

Random Forest is used for both Regression and Classifier problems. It is used to divide the main tree in to sub trees randomly. The more the number of sub tree the accurate the result. (SVM) Support Vector Machine algorithm is also used for both the regression and classification problems. It is a supervised ML Algorithm.

4.1 System Architecture:



This is the basic architecture of my project where I going to implement the project on the main server and the host it in the local host using the flask server and once the application is up and running on the localhost. Keep note of the port Id and by using the ngrok proxy server which is going to host the web application publically by forwarding the web application using the port id on internet & we can access it using the Domain Name generated by the ngrok from anywhere.

4.2 Dataset Description

- **Address bar based features**: these include the Using of IP Addresses, Increasing the Length of the URL to Hide the Suspicious Path, using 3rd party services and Shortening the URL, URL with Special Symbol ‘ @

’, using ‘ // ’ for Redirecting, using ‘ - ’ to add Prefix and Suffix in the Domain, creating multiple subdomains, HTTPS, Domain-Registration length, Favicon, Utilizing Non Standard Ports, Appearance of ‘HTTPS’ Token.

- **Abnormal Based Features:** Theses include URL Requests, Anchor URLs, Links Present in meta tag, script tag, SFH, Notification to Email Regarding the submission,
- **HTML and JavaScript based Features:** Forwarding Websites, Customizing Status Bar, Right click Disable, Utilizing Pop-up’s, Redirecting Iframe.
- **Domain based Features:** Domain Age, Domain Name System Record, Traffic in website, web page rank, google index, list of links pointing to the web page.

5. METHODOLOGY

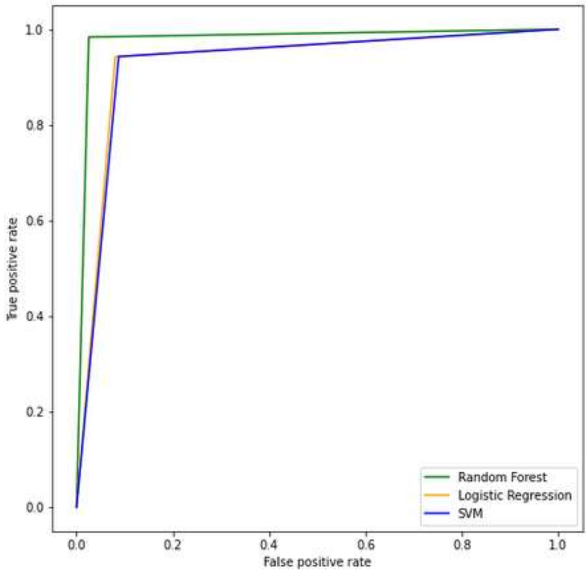
Here we are going to study all the features mentioned above to predict the accuracy of the website wheatear it is legitimate or not. This function can be performed using the algorithms like Logistic Regression, Random Forest, Support Vector Machine (SVM) Algorithm. The Algorithm is been Triggered whenever the user enters the URL. The part of the algorithm is to retrieve the features of the site utilizing a URL. The URL is used to retrieve the URLs and page rank's features and so forth. the retrieved features will be shipped off the classifier to deliver the objective name that shows the condition of the site at that point executes the appropriate activity on that.

We have built the model using Random Forest Technique, Linear Regression algorithm & Support Vector Machine Algorithm which follows as mentioned below. First, we are going load the dataset & split the data into train dataset and test dataset. In the ratio of 20% for Testing the dataset & 80% for Training the dataset.

Here we are going to train and test all combinations of features mentioned in the dataset description section to find the accuracy of the features. Once the feature extraction is done, we are going to check the rank of the website and based on the accuracy. & decide whether the URL is a Phishing web site or a Legitimate web site.

I have used the ngrok which is used to deploy the localhost web page publically. Basically, it is tool which is uses to for port forwarding by using this tool I was able to make my phishing web application available online publically through my private server or laptop.

6. RESULT ANALYSIS



After implementing the Logistic Regression Random Forest Algorithm & support vector machine Algorithm (SVM) below is Comparison chart what I have got as the accuracy. Comparison of Logistic Regression, Random Forest & SVM Algorithm. We can see the all the Models ae performing good because all the models are giving the accuracy of more than 90% consistently, but we can clearly see that the Random Forest Algorithm has got the most Accuracy like above 96% - 97% most of the time and the second best algorithm is Support Vector Machine Algorithm (SVM) which has the accuracy levels between 93% - 95%, and Finally the Logistic Regression is also good but when compared to the Random Forest and Support Vector Machine Algorithm it has scored a bit less accuracy between 91% - 92%. Hence the project was successfully implemented.

7. CONCLUSION

URLS	**Logistic *Regression***	***Random Forest***	***SVM***
https://www.facebook.com/	92.85	97.15	94.43
https://www.linkedin.com/	92.89	96.92	94.84
https://www.youtube.com/	92.53	97.37	94.34
https://www.javatpoint.com/	92.94	97.10	93.66
https://www.tutorialspoint.com/index.htm	92.80	96.92	93.89
https://www.flipkart.com/	92.80	97.19	94.21
https://www.google.com/	92.58	96.74	94.79
http://encouragings.ml/section/image/file/outlook/ad7173b3253253611fecf452e4e00d05/	92.62	97.06	95.61
http://datarescue.cl/bace/907c20172c04ac2b303185acdd47ed3e?login=&.verify%3Fservice=mail&data:text%2Fhtml%3Bcharset=utf-8%3Bbase64%2CPGh0bWw%2BDQo8c3R5bGU%2BIGJvZHkgeyBtYXJnaW46IDA7IG92ZXJmbG93OiBoaWRkZW47IH0gPC9zdHlsZT4NCiAgPGlmcmFt	91.81	97.64	95.74
http://127.0.0.1:5000/	92.35	97.24	95.02

After exploring and reviewing for suitable monitoring tools, proposed framework has been recognized and decided to address the complexity of monitoring requirement for current circumstance. This proposed software is intended to show awareness to the broad level of its functionalities, features that can be shown in this Monitoring Era. The Proposed framework prevents from significant data leak-out, produce better control component and alerts the client to guard their private data. Like some other projects, there are upgrades which could be made into this framework in future.

8. REFERENCES

[1]**Author:** Almaha Abuzuraiq, Mouhammd Alkasassbeh, and Mohammad Almseidin, “Intelligent Methods for Accurately Detecting
Phishing Websites”. 2020 *IEEE 10.1109/ICICS49469.2020.239509*.
[2]**Author:** R. Kiruthiga, D. Akila “Phishing Websites Detection Using Machine Learning,” *ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019 (IJRTE).*
[3]**Author:** Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, Dr.Aram Alsedrani “Detecting Phishing Websites Using Machine Learning,” *978-1-7281-0108-8/19/* ***Date: 2019*** *IEEE*
[4]**Author:** Mohammed Hazim Alkawaz, Stephanie Joanne Steven and Asif Iqbal Hajamydeen, “Detecting Phishing Website Using Machine Learning,” *978-1-7281-5310-0/20/* ***Date: feb. 2020*** *IEEE*
[5]**Author:** Arun Kulkarni 1 , Leonard L. Brown, “Phishing Websites Detection using Machine Learning” Vol. 10, No. 7, 2019 IJACSA
[6]**Author:** Shraddha Parekh, Dhwanil Parikh, Srushti Kotak, Smita Sankhe, “A new method for Detection of Phishing Websites: URL Detection”,
IEEE Xplore Compliant 2018 - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2
[7]**Author:** Rishikesh Mahajan, Irfan Siddavatam, “Phishing Website Detection using Machine Learning Algorithms”, International Journal of Computer Applications (0975 – 8887) Volume 181 – No. 23, October 2018
[8]**Author:** SU Yang, “Research on Website Phishing Detection Based on LSTM RNN”, *978-1-7281-4390-3/20 Date: 2020 IEEE*
[9]**Author:** Mehek Thaker, Mihir Parikh , Preetika Shetty , Vinit Neogi , Shree Jaswal, “Detecting Phishing Websites using Data Mining”, *IEEE Xplore 2018 ISBN:978-1-5386-0965-1*
[10]**Author:** Huaping Yuan, Xu Chen, Yukun Li, Zhenguo Yang, Wenyin Liu**,** “Detecting Phishing Websites and Targets Based on URLs and Webpage Links” *IEEE 978-1-5386-3788-3/18 August 20-24, 2018*