International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 5 Issue 1, November-December 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Phishing URL Detection

Dirash A R[1], Mehtab Mehdi[2]

[1]Student, [2]Assistant Professor,
[1,2]Jain Deemed-to-be University, Bengaluru, Karnataka, India

ABSTRACT

Phishing is a method of trying to gather personal information using deceptive emails and website; it is a classic example for cybercrime. For example we may receive an email from our bank or trusted company and its asks you for information which may look real but it's designed to fool you into handing over crucial information this is a scam and we need to avoid it. There are many techniques to detect it but Machine learning is the most effective technique for detecting these types of attacks and it can detect the drawbacks of other phishing techniques. This paper focuses on discerning the many features that discriminate between authorized and phishing URLs. The main aim of this paper is to develop a model as a solution for detecting malicious websites. By detecting a large number of phishing hosts, this model can manage 80-95 percent accuracy while retaining a modest false positive rate. Implementation will be carried out on the datasets of 4,20,465 websites containing both phi shy sites and authorized sites. Ultimately, the findings will show us the higher precision detection rate algorithm, which will classify phishing or legitimate websites more correctly.

KEYWORD: *Malicious Identification, Malicious website, Logistic regression, confusion matrix*

How to cite this paper: Dirash A R | Mehtab Mehdi "Phishing URL Detection" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-5 | Issue-1, December 2020, pp.980-982, URL: www.ijtsrd.com/papers/ijtsrd38109.pdf

IJTSRD38109

INTRODUCTION

A "phishing" or "spoofed" website is a website that looks like a legitimate website Phishing is misdirecting you to associations to take your sensitive details. It might be an email that looks like it came from a bank, or it might be a connection that seems to compel you to sign in to your account again. In this scenario, the sender would have access to your accounts and important information by sending you to a site or connection and sharing your account details. Phishing is usually achieved by spoofing emails or instant messages, and sometimes directs users to enter information on a fake website whose appearance and feel are almost identical to the real one. There are a variety of different methods that are used to get users personal information. As technology advances, the tactics employed by cybercriminals are becoming more sophisticated.

Phising attacks are very easy to set up so you need to be very careful when providing information on any of the websites. For a phising attack hackers create a duplicate page on any website and set up the page with the support of social engineering to pass the data you enter on the site to the hackers host. So, they can get all your account information, including your username, password, credit card details, etc.

The purpose of this project is to train machine learning models using logistic regression in a dataset designed to predict phishing websites. Both phishing and duplicate website URLs are collected to form a dataset and the necessary URL and content-based features of the website are extracted from them. The performance level of the model is calculated here.

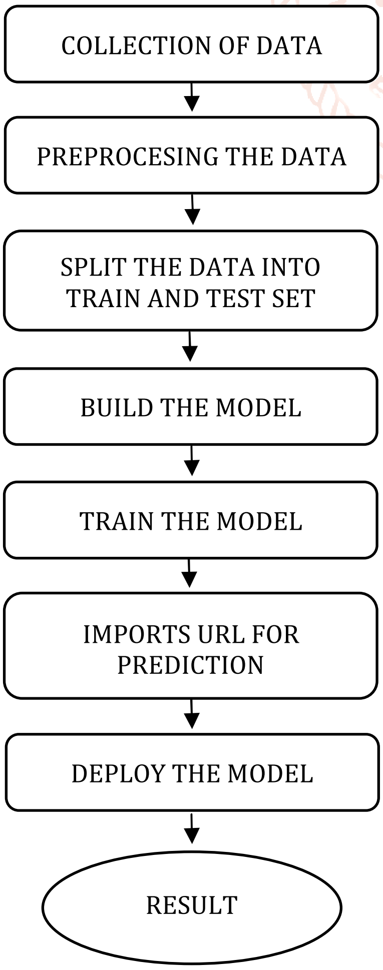
Literature Review:

The table below shows what people have contributed to identifying or detecting the malicious in URL.

AUTHOUR	YEAR	DISCRIPTION
Seiferte et al[1]	2008	This paper introduces a novel classification method for detecting malicious web pages, which includes inspecting the underlying server relationships
Justin and Saul et al [2]	2009	In this paper we present an approach to this problem based on automated URL classification, using statistical methods to detect the tell-tale lexical and host-related properties of malicious Web site URLs
Hou et al [3]	2010	In this paper we present an approach to this problem based on automated URL classification, using statistical methods to detect the tell-tale lexical and host-related properties of malicious Web site URLs
Welch et al[4]	2011	This paper introduces a novel two-stage model for the identification of malicious web pages
Choi et al[5]	2011	In this paper, they propose method using machine learning to detect suspected URLs of all the popular attack types and identify their nature.
Sirageldin et al[6]	2014	In this paper, they provide a framework for detecting a suspected website using artificial neural network learning technique
Michael and Heileman et al[7]	2015	They presented in this paper a lightweight method for identifying malicious web pages using lexical URL analysis alone
Hu et al[8]	2016	They concluded that machine learning techniques can accurately classify malicious domains in various ways. The BPSO feature also increased the performance
Desai et al[9]	2017	They Built a extension for Chrome that acts as a middleware between users and malicious websites, and reduce the possibility that users may succumb to such websites
Sahingoz et al[10]	2018	This paper proposes a real-time anti-phishing scheme, which uses seven different classification algorithms and features based on natural language processing (NLP)
Alkawaz et al[11]	2020	This paper proposes a phishing detection system approach for detecting blacklisted URLs, so that it will alert everyone while browsing or accessing a specific website
Singh et al[13]	2020	The analysis raises awareness of phishing attacks, detects phishing attacks and encourages the readers to practice phishing prevention

METHODOLOGY:

Phishing websites can't be easily identified, so the machine learning solution is the perfect way to find them. The key benefit of machine learning is its self-learning capability. We use the urldata dataset to train the model. The key steps involved in the implementation are:



COLLECTION OF DATASET

The dataset is collected from the PHISHTANK website, which includes both phishy and authorized sites. This dataset consists of 4,20,465 entries of websites categorized as good and bad and it contains different varieties of websites.

PREPROCESING THE DATA

Data Preprocessing is the cycle wherein we make the information reasonable to be performed over Model with less exertion. A URL is made up of some significant or meaningless words and some special characters which should be removed and also remove the words that are repeated. They also remove special characters like the '/' and '-' which can be done using TfidfVectorizer.

SPLITTING THE DATA INTO TRAIN AND TEST SET

Here we split the data into train and test set in the ratio 80:20.The training and Testing is done for model validation, which helps in building the model. Out of all the data, 80 percent of the population will be trained to build the model. Then you can verify your model by making the predictions for the remaining observations called test set.

BUILD THE MODEL

We build the model using the training set and also using the logistic regression algorithm. The logistic regression algorithm is a built in function so we don't need to implement it, we just need to call the function.

DEPLOY THE MODEL

After training and testing the model, the next step is to deploy the model and check the accuracy of the model being deployed. The performance of the model is checked through confusion matrix. The real values and the predicted values are tabulated in a 2×2 matrix.

Result and Discussion

The Model is designed to detect malicious URLs using Logistic Regression algorithm. Strategic Regression will give an exactness of over 90% and it isn't hard to implement when contrasted with different models. After undergoing all the steps, we add some URLs that's good and fake, and our model predicts which URLs is good and which is bad.

Conclusion and Future Work

Conclusion:

There are a variety of machine learning applications in computer and network security, and the concept of detecting malicious URLs is a key one. Phising attacks are very easy to set up so you need to be very careful when providing information on any of the websites.

There are several algorithm for detecting the malicious URL, but logistic regression is better compared to other algorithm, as it is easy to implement. By Fitting logistic regression and creating confusion matrix of predicted values and real values the above model was able to get around 96% accuracy

Future Work:

Through this project, one will know a lot about phishing websites and how they are differentiated from legitimate websites. For future work, we expect to improve our model by trying algorithms other than logistic regression, developing new and improved features, and also trying out a training model using mapped features.

This project can be taken further by creating a browser extensions of developing a GUI. There will be several advanced models in the future and their predictions will be more reliable, resulting in better prediction.

References

[1] C. a. o. Singh, "Phishing Website Detection Based on Machine Learning: A Survey," 2020.

[2] M. H. a. S. S. J. a. H. A. I. Alkawaz, "Detecting Phishing Website Using Machine Learning," 2020.

[3] C. a. W. I. a. K. P. a. A. C. U. a. E.-P. B. Seifert, "Identification of malicious web pages through analysis of underlying dns and web server relationships," 2008.

[4] J. a. S. L. K. a. S. S. a. V. G. M. Ma, "Beyond blacklists: learning to detect malicious web sites from suspicious URLs," 2009.

[5] Y.-T. a. C. Y. a. C. T. a. L. C.-S. a. C. C.-M. Hou, "Malicious web content detection by machine learning," 2010.

[6] I. a. G. X. a. K. P. a. o. Welch, "Two-stage classification model to detect malicious web pages," 2011.

[7] H. a. Z. B. B. a. L. H. Choi, "Detecting Malicious Web Links and Identifying Their Attack Types," 2011.

[8] A. a. B. B. B. a. J. L. T. Sirageldin, "Malicious web page detection: A machine learning approach," 2014.

[9] M. a. H. G. a. G. G. a. A. A. a. P. P. Darling, "A lexical approach for classifying malicious URLs," 2015.

[10] Z. a. C. R. a. P. I. a. S. W. a. B. Y. Hu, "Identifying malicious web domains using machine learning techniques with online credibility and performance data," 2016.

[11] A. a. J. J. a. N. R. a. R. N. Desai, "Malicious web content detection using machine leaning," in 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information \& Communication Technology (RTEICT, 2017.

[12] O. K. a. B. E. a. D. O. a. D. B. Sahingoz, "Machine learning based phishing detection from URL," 2019.

[13] M. H. a. S. S. J. a. H. A. I. Alkawaz, "Detecting Phishing Website Using Machine Learning," 2020.