International Journal of Modern Developments in Engineering and Science
Volume 1, Issue 5, May 2022
https://www.ijmdes.com | ISSN (Online): 2583-3138

A Review on Cyber Threats Analysis Using Data Mining Techniques – With Special Reference to Phishing Attacks

M. Rajitha[1*], R. Priya[2]

[1]*Research Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, India*
[2]*Associate Professor, Department of Computer Science, Sree Narayana Guru College, Coimbatore, India*

***Abstract*: The article discusses the reviews and problem of effective monitoring of measures taken to enforce cyber security. Phishing is the most commonly used social engineering and cyber-attack. Through such attacks, the phisher primarily targets typical online users by tricking them into revealing confidential information, with the purpose of using it for fraud. To avoid getting phished, users should be aware of phishing websites. To form blacklist of phishing websites, it requires information of website being detected as phishing. Detect them in their early appearance, using Machine Learning and Deep Neural Network algorithms. The machine learning based method is proven to be most effective and useful than the other methods. Albeit, online users are still being trapped or forced to reveal sensitive information in phishing websites.**

***Keywords*: Cyber threats, Phishing, Data mining, Machine Learning.**

1. Introduction

A phishing website is a common social engineering method tool that looks like trustful Uniform Resource Locators (URLs) and WebPages. The objective of this work is to train machine learning models and deep neural nets on the dataset created to predict phishing websites. phishing and legitimate URLs of websites are gathered to form a dataset and from these the required URL features and website content-based characteristics are extracted. The performance level of each model will be measured, analyzed and compared.

2. Conceptual Design

Data Mining: Data mining allows users to sift through the enormous amount of information available in data warehouses and extracting the necessary knowledge which can be used for the decision-making purpose. Data Mining allows backend processors to analyze data from many different dimensions, categories it and summarize the relationships identified. Majority of organizations in the field of business are using Data Mining in one-way or the other. Jyothi Pillai in “User centric approach to item set utility mining in Market Basket Analysis” Explains the importance of Business intelligence and is the information about a company's history and old performance that is used to predict the future performance. Association rule mining is one of the techniques used an in data mining research where the aim is to find interesting correlations among sets of items in databases.

Cyber Threat: Cyber threat or cyber security threat is a malicious act that seeks to damage data, steal data, or disrupt digital life in general. These include various types of threats like computer viruses, data breaches, Denial of Service (DoS) attacks, and other attack vectors. It also refers to the possibility of happening cyber-attack that aims to gain unauthorized access, damage, disrupt, or steal an information technology asset, computer network, intellectual property or any other form of confidential data. Cyber threats can be sourced from within an organization by trusted users or from anonymous locations by unknown persons.

Phishing: Phishing is the most common type of social engineering attack frequently used to steal user data, including login credentials and credit card information. It happens when an attacker, behaves like a trusted entity, dupes a victim into open an email, instant message, or text message. The person is then forced to click a particular malicious link, which may lead to the installation of malware programs, or it leads to revealing of sensitive information. An attack can have devastating results. For persons, this includes non-aware purchases, stealing of funds, or identifies theft.

Moreover, phishing is often used to attain confidential information in corporate or governmental networks as a part of a larger attack, such as an Advanced Persistent Threat (APT) event. So that employees are compromised in order to bypass security parameters and distribute malware within a closed environment, or can privileged access to secured data.

An organization affected to such an attack typically sustains severe financial and economic losses in addition to declining market share, reputation, and consumer trust. Depending on scope and wide, a phishing attempt may elaborate into a security incident from which a business will have a difficult time recovering.

*Corresponding author: vanirajitha07@gmail.com

3. Approach

Below mentioned are the steps involved in the completion of this work project: Collect dataset containing phishing and legitimate websites from the open-source platforms. Next is to write a code to extract the required features from the URL database. The extracted features are analyzed and preprocess the dataset by using various EDA techniques. Divide the dataset into training and testing sets. Run selected machine learning and deep neural network algorithms like SVM, Random Forest, KNN, XGBoost on the dataset. The next step is to Write code for displaying the calculated result considering accuracy measures. Compare the obtained results for trained models and specify which is better.

4. Data Collection

Valid URLs are collected from the dataset maintained by University of New Brunswick. From this collection, randomly selected 5000 URLs. Phishing URLs are collected from open-source service called Phish Tank. It provides a set of phishing URLs in heterogeneous formats like csv, json etc. that gets updated hourly. Form the obtained collection, 5000 URLs are randomly picked.

5. Machine Learning Models

This is a supervised machine learning task. The major classification of supervised machine learning methods, called classification and regression. This data set comes under classification problem, as the input URL is classified as phishing (1) or legitimate (0). The machine learning models in classification, which are selected to train the dataset in this are Random Forest, XGBoost, KNN, Support Vector Machines.

6. Findings

- Phishing sites can be identified and precautions can be taken.
- It is very useful to identify which algorithm gives accurate results in prediction.
- The performance level of each model is measured and compared to gain efficient results.
- We can use the analyzed data to keep track of vulnerabilities in the system.
- Data sets can be elaborated using training in the cases of new threats happening.
- We can make use of the efficient algorithm to predict whether a URL is malicious or not.

7. Conclusions and Future Scope

To conclude that it is possible to use data mining techniques like classification and prediction as well as machine learning algorithms to predict a given URL is a cyber threat or not. In recent years, with the increasing usage of mobile technologies, there is a trend to move almost all real-world tasks in to the cyber world. Although this makes our daily lives easy, it also brings many security breaches due to the anonymous structure of the Internet. Used antivirus programs and firewall systems can prevent most of the attacks. However, experienced attackers target on the weakness of the computer users by trying to phish them with bogus WebPages. These pages look like some popular or frequently used banking, social media, e-commerce sites etc. sites to steal some sensitive information such as, user-ids, passwords, bank account, credit card numbers, etc. Phishing detection is a challenging problem. Further development of such a system can be continued by adding statistical tools and machine learning tools to obtain additional information about threats on their own, which will reduce the requirements for data sources and allow you to expand the set of analyzed indicators.

References

[1] S. Vijaylakshmi, V. Mohan, S. Suresh Raja, “Mining of users access behavior for frequent sequential pattern from web logs” International Journal of Database Management System, Vol. 2, August 2010.

[2] Edi Winarko and John F. Roddick, “Discovering Richer Temporal Association Rules from Interval-Based Data, Data Warehousing and Knowledge Discovery,” 2005.

[3] Shrivastava A., and Sahu R., “Efficient Association Rule Mining for Market Basket Analysis,” Global Journal of e-Business & Knowledge Management, Volume 3, Issue 1, 2007.

[4] Aggarwal C.C, “Mining association with the collective strength approach”, 2001.

[5] Jianying Hu, Aleksandra Mojsilovic, “High-utility pattern mining: A method for discovery of high-utility item sets,” Pattern Recognition, vol. 40, no. 11, November 2007.

[6] N. R. Srinivasa Raghavan. “Data Mining in E-commerce: A Survey”. Sadhana, vol. 30, no. 2, 2005.

[7] Mobasher B., “Web Usage Mining and Personalization,” Practical Handbook of Internet Computing (ed.) M. P. Singh (CRC Press), 2004.

[8] Jiauei Han, Michele Kamber, Simon Fraser, “Data mining Concepts and Techniques,” ISBN 1-55860-489-8-2001.

[9] Randall S. Sexton, Richard A. and Michael A., “Predicting Internet/e-commerce use”, Internet Research, vol. 5, 2002.

[10] J. Hu, and A. Mojsilovic, “High-utility pattern mining: A method for discovery of high-utility item sets”, Pattern Recognition, 2007.

[11] https://www.stealthlabs.com/blog/cyber-security-threats-all-you-need-to-know/

[12] https://www.microstrategy.com/us/resources/introductory-guides/data-mining-explained

[13] https://www.upguard.com/blog/cyber-threathttps://www.upgrad.com/blog/association-rule-mining-an-overview-and-its-applications

[14] https://towardsdatascience.com/association-rules-2-aa9a77241654