WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Detecting Phishing Websites Using Machine Learning Technique

Kshitij Bohre, Kishore sohanda, Kunal patel, Harshit saxena

Department of Computer Science and Engineering, Acropolis Institute of Technology and Research, Indore, Madhya Pradesh, India.

ABSTRACT

Phishing attacks are the easiest way to obtain sensitive informationfrom innocent users. The aim of the phisher is to obtain his vital information such as username, password, bank account details, etc. Cybersecurity professionals are currently looking for reliable and trustworthy detection techniques to detect phishing websites.This white paper extracts various characteristics of legitimate and phishing URLs. Learn about machine learning technology that detects phishing URLs by analyzing Decision Trees, Random Forests, and Support Vector Machines .The algorithm is used to detect phishing websites. The purpose of the paper is to narrow down the best machine learning algorithms by detecting phishing URLs and comparing the accuracy, false positive, and false negative rates of each algorithm.

Key-Words: - Phishing attack, Machine learning, Cyber Security.

Introduction

Phishing is now a major concern ofsecurity researchers. This is because it is not difficult to create a fake website that closely resembles a legitimate website. Expertscan identify fake websites, but not all users can identify fake her websites and such users become victims of her phishing attacks. The attacker's main goal is to steal Bank's credentials. A US company loses $2 billion a year to her
Because customers are victims of her phishing [1]. The 3rd Microsoft Computing Safer Index Report, published in February 2014, estimated that the annual global impact of phishing could reach $5 billion [2]. Phishing attacks succeed because has no user awareness. Phishing attacks exploit user vulnerabilities discovered in, making them very difficult to mitigate, but improving detection techniques for phishing is very important.

Problem Formulation

Phishing detection technology suffers from low detection accuracy and high false positives, especially when new phishing techniques are introduced. Also, the most commonly used blacklist-based techniques are inefficient in responding to outbound phishing attacks. This is because it makes it easier to register new domains and there is no comprehensive blacklist that can guarantee a completely up-to-date database. In addition, page content checks are used in several strategies to overcome the problem of false negatives and compensate for vulnerabilities in older lists. Additionally, the algorithms used to check page content have different approaches and varying degrees of accuracy for detecting phishing websites. Therefore, ensembles can be viewed as better solutions because they can combine similarity of accuracy and different error coverage characteristics in the algorithm of choice. Therefore, this study addresses several findings:
1.How is the raw dataset processed for phishing detection?
2. How does the phishing website's algorithm increase detection rates?
3. How to reduce the false negative rate in the phishing website algorithm

Literature Review

Phishing attacks are classified according to the phisher mechanism used to catch suspicious users. Some forms of these attacks are key loggers, DNS poisoning, etc. [2]. The social engineering initiation process includes online blogs, short messaging (SMS) services, social media platforms using Web 2.0 services such as Facebook and Twitter, peer file sharing services, and Voice-over-IP (VoIP) systems. , the attacker uses the spoofed identity of the caller. [3, 4]. To trick unsuspecting consumers, each form of phishing performs the process slightly differently. An email phishing attack occurs when an attacker sends an email containing a link to a potential user to redirect the user to a phishing website.

Methodology

Three machine learning classification model Decision Tree,Random forest and Support vector machine has been selected to detect phishing websites.

1. Decision tree algorithm

One of the most used algorithms in machine learning technique. The decision tree algorithm is easy to understand andalso easy to implement. The decision tree begins its workselecting the best allocator from the available attributes fora classification that is considered the root of the tree.The algorithm continues building the tree until it finds a leaf node.A decision tree creates a training model that is used for predictiontarget value or class in the tree representation of each internal nodethe attribute and each leaf node of the tree belong to the treebelongs to the class designation. In the decision tree algorithm, the gini indexand information retrieval methods are used to calculate them
nodes.

2.Random Forest Algorithm

The random forest algorithm is one of the most powerful algorithms in machine learning technology and is based on decision tree algorithm concept. Random forest algorithm creates a forest with many decision trees. High number tree provides high detection accuracy. The creation of trees is based on the bootstrap method. In bootstrap method properties and dataset samples are randomly selected with compensation for creating a single tree. Between randomly selected elements, a random forest algorithm selects the best
Classification splitter and similar decision tree algorithm; the random forest algorithm also uses the gini index and information profit methods to find the best distributor. This process will get continue until the random forest produces n number of trees. Each tree in the forest predicts the target value and then the algorithm tallies the votes for each predicted target. Finally the random forest algorithm takes into account the predicted target with a high vote as a final prediction.

3.Support Vector Machine Algorithm

Support vector machine is another powerful algorithm machine learning technology. In support vector machine algorithm each data item is plotted as a point in n-dimension spatial and support vector machine algorithms dividing line for the classification of two classes, this separation
the line is well known as the hyperplane. Support vector machine looks for the nearest points called as support vectors and once it finds the closest point it draws a line that connects to them. Then support vector machine construct a dividing line that bisects and is perpendicular to it
connecting line. So that the data can be perfectly classified margin should be maximum. Here is the edge distance between the hyperplane and the support vectors. In a real world scenario it is
it is not possible to separate complex and non-linear data, solve this problem is supported by a vector machine using a kernel trick that transforms a lower dimensional space into a higher dimensionspace.

Result Discussions

- The model gives us an alert regarding suspicious website.
- The model classifies the website being used into phishing website or safe to use website.
- New internet users will easily be able to browse internet safely.

Limitations of the Work

- One of the Limitations can be thatour model is able to detect website but not prevent one from again using them again.
- Model is highly dependent on accuracy of features.
- Machine learning based systems cannot correctly utilize such a large dataset.
- Classifier may not do well with large datasets.

Suggestion and Recommendations for Future Work

- In the future we can expand our model to not only detect but prevent one from using it again.
- A chrome extension can be made using this model so that one can easily add it to google chrome without any additional installations.

Conclusion

The proposed study emphasized the phishing technique in the context of classification wherePhishing websites are believed to involve automatically categorizing websites into a predetermined set of class values based on several features and a class variable. ML-based phishing techniques depend on website features to gather information that can help classify those websites to detect phishing sites. The problem of phishing cannot be eradicated, but it

can be reduced by fighting it in two ways, by improving targeted anti-phishing practices and techniques and informing the public about how fraudulent phishing sites can be detected and identified to combat the ever-evolving and sophisticated phishing attacks and tactics, ML anti-phishing techniques are essential. The authors used the LSTM technique for identification malicious and legitimate websites. A crawler was developed that crawled 7900 URLs Alexa-Rank portal and also used the Phish-tank dataset to measure the effectiveness of the proposed URL detector. The result of this study shows that the proposed method represents better results than existing deep learning methods. A total of 7900 malicious The URLs were detected using the proposed URL detector. This resulted in better accuracy and F1 - Timed Score. The future direction of this study is to develop an unsupervised deep learning method for generating statistics from a URL. In addition, the study can be expanded to create a result for the larger network and protect the company's privacy individual.

Acknowledgment

We thank the almighty Lord for giving us the strength and courage to sail out through the tough and reach on shore safely.

There are number of people without whom this projects work would not have been feasible. Their high academic standards and personal integrity provided me with continuous guidance and support.

We owe a debt of sincere gratitude, deep sense of reverence and respect to our guide and mentor**Priti shukla,** Professor, AITR, and**Praveen bhanodia, Professor, AITR**, Indore for their motivation, sagacious guidance, constant encouragement, vigilant supervision and valuable critical appreciation throughout this project work, which helped us to successfully complete the project on time.

We express profound gratitude and heartfelt thanks to **Dr. Kamal Kumar Sethi,** HOD, CSE, AITR Indore for his support, suggestion and inspiration for carrying out this project. I am very much thankful to other faculty and staff members of CSE Dept., AITR Indore for providing me all support, help and advice during the project. We would be failing in our duty if do not acknowledge the support and guidance received from **Dr. SC Sharma**, Director, AITR, Indore whenever needed. We take opportunity to convey my regards to the management of Acropolis Institute, Indore for extending academic and administrative support and providing me all necessary facilities for project to achieve our objectives.

We are grateful to **our parents** and **family members** who have always loved and supported us unconditionally. To all of them, we want to say "Thank you", for being the best family that one could ever have and without whom none of this would have been possible.

References

[1] Pujara, P., &Chaudhari, M. B. (2018). Phishing website detection using machine learning: a review. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 3(7), 395-399.

[2] Mahajan, R., &Siddavatam, I. (2018). Phishing website detection using machine learning algorithms. International Journal of Computer Applications, 181(23), 45-47.

[3] Kulkarni, A. D., & Brown III, L. L. (2019). Phishing websites detection using machine learning.

[4] Zamir, A., Khan, H. U., Iqbal, T., Yousaf, N., Aslam, F., Anjum, A., &Hamdani, M. (2020). Phishing web site detection using diverse machine learning algorithms. The Electronic Library, 38(1), 65-80.

[5]Singh, C. (2020, March). Phishing website detection based on machine learning: A survey. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 398-404). IEEE.