www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 04, April 2023, pp : 515-518

e-ISSN : 2583-1062

Impact Factor : 5.725

PHISHING WEBSITE DETECTION USING MACHINE LEARNING

Pavan Suhas Reddy Medapati[1], Pyla SriLekha[2], Mulakala DivyaSree Yadav[3], N. Siva Kumar[4]

[1,2,3]CSE, B. Tech Aditya Engineering College Surampalem

[4]Guide: M.Tech., (Ph. D) Sr. Assistant Professor Aditya Engineering College Surampalem

ABSTRACT

The most dangerous criminal activity in cyberspace is phishing. The huge number of populated people access administrative and financial institution services online, phishing attacks have rapidly developed over the past decades. The first time scammers made money, they turned it into a lucrative business. Phishing is a tactic used frequently to trick people into providing their sensitive content by implementing phoney websites. Phishing website domains are used to raid sensitive information like userid's, countersigns, and other bank content. Phishers use web browsers that aesthetically and linguistically resemble the genuine websites. Phishing techniques started to advance quickly as technology continued to advance, and this should to be terminated by using anti-social techniques to identify unsafe browsers. To combat phishing attacks, machine learning is a potent tool. This study examines machine learning-based detection methods and the features that go into those methods.

Keywords –AdaBoost, Random forest, XGBoost, performance Analysis, Gradient Boosting and support vector machine

1. INTRODUCTION

The most threatful criminal business in cyberspace is phishing. The huge number of populated people access administrative and financial institution services online, phishing attacks have rapidly developed over the past decades. Spoofers began to generate money and now leading a successful business. Spoofers threat innocent users using a variety of methodologies, including messaging, spoofed web browsers, and fake domains. It is very easy to create fake domains that, in terms of design and content, resemble real websites. Additionally, attackers pose as risky-level security precautions and provide subscribers with high-level security questions and answers. Subscribers who answer those questions are more open to spoofing scams. Numerous studies have been initiated to terminate fraudulent attacks by various organizations globally. By recognizing the domains and training subscribers to detect the fraud domains, phishing attacks can be suspended. One of the effective methods for spotting phishing websites has been machine learning methodologies. This study has discussed a number of techniques for spotting fraudulent web pages. Machine learning is a branch of AI that concentrates on creating algos and mathematical equations that let computers automatically get better at a particular task over time. By building predictive models that can be trained on data, machine learning aims to produce predictions or decisions that are accurate when applied to fresh data.

2. LITERATURE REVIEW

This article conducts a review of the literature on phishing attack detection. Phishing attacks aim to exploit holes left by the human factor in systems. Users are the weakest link in the security chain because a large number of cyberattacks are spread via mechanisms that take advantage of user vulnerabilities. Since there is no one particular explanation to effectively address all the risk factors in spoofing, multiple methodologies were frequently used to counteract particular attacks. The purpose of this paper is to review a number of the recently introduced phishing lessening techniques. We believe that it is essential to demonstrate the phishing recognition methodologies fit into the overall lessening process to provide a high-profile summary of the different categories of phishing deduction techniques. [1]A significant rise in electronic trading, or consumer-to-consumer online transactions, has occurred recently as a result of advances in Internet and cloud technologies. The resources of an enterprise are harmed by this growth, which allows unauthorised access to sensitive user data. One of the well-known methods of tricking users into accessing harmful content and giving up their information is phishing. Most phishing pages have exact replicas of legitimate browsers in terms of their webpage UI and domain URL's. Blacklist, heuristic, and other detection methods for phishing websites have all been proposed. There is, however, a sharp rise in the number of victims as a result of inadequate security measures. Phishing attacks are more likely to succeed online because of its unregulated and anonymous nature. [2]

3. PROPOSED METHODOLOGY

In the modern world, machine learning is cutting edge and popular for a wide range of applications because it can handle massive amounts of data, has been updated with new algorithms, and has powerful GPU processing power. Four machine learning algorithms are used in this project to handle these massive data sets from each data set,

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 04, April 2023, pp : 515-518

e-ISSN : 2583-1062

Impact Factor : 5.725

maximising usage and sending accurate reports.Algorithms Used: ADABoost Classifier, XGBoost, Random Forest Classifier, Gradient Boosting Classifier.These algorithms will not affect the performance of the local machine and handles the huge data sets such that it opitmizes the space as well as time consumed. Accurate results are obtained and are being sent to the cyber security team as well.

Hardware requirements:

- Windows 7 or 7+
- Hard Disk: 500GB
- RAM - 8GB
- Processor used: Intel 3rd gen

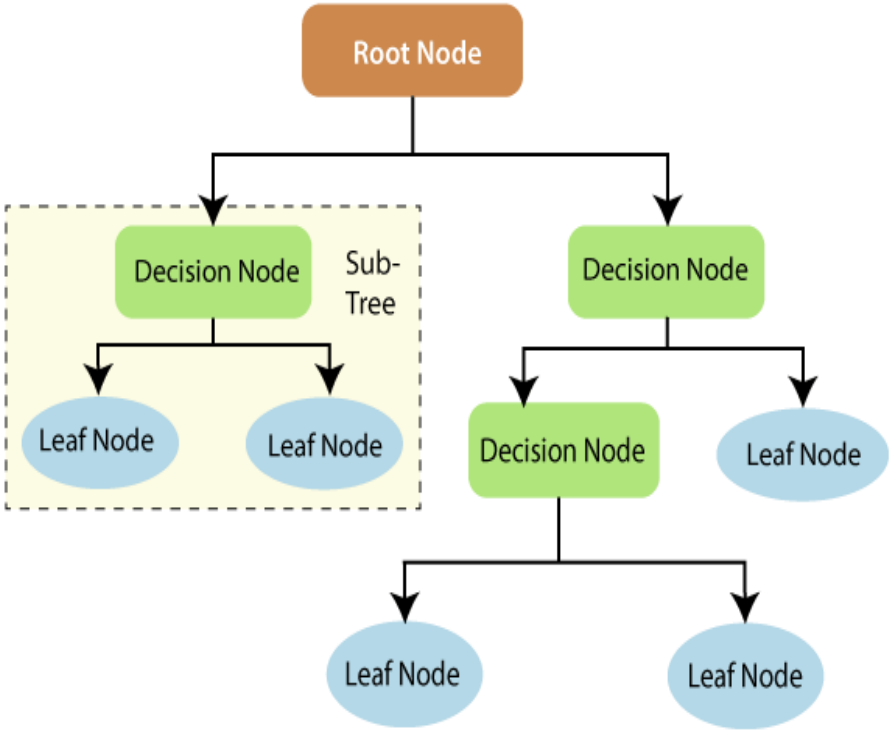
Software requirements:

- Software: Python 3.6
- PyCharm
- FLASK
- Libraries Used – Pandas, urllib, sklearn, os. Ipaddress

4. ALGORITHMS USED

Random Forest Classifier:

To address issues with regression and classification, a random forest machine learning method is used. It employs collaborative learning, an approach which utilises multiple types of classifiers to provide answers to difficult questions. Random forest algorithms are made up of deterministic trees. The random forest method trains the "forest" it creates by using bagging or bootstrap aggregation. The accuracy of machine learning algorithms is improved by the bagging process, an ensemble meta-algorithm. The RFC algorithm calculates the result related on the predictions of the decision trees. The output of various trees has been averaged or averaged when making projections. While there are more trees, the outcome is more accurately predicted. This algorithm terminates the drawbacks of the decision tree algorithm.



.

XGBoost Algorithm:

The term "Extreme Gradient Boosting" refers to this process. A distributed gradient boosting library that has been optimised for speed, adaptability, and portability is called XGBoost. Gradient Boosting is used to implement machine learning algorithms. Boosters are a wide technique to ensemble learning for constructing a powerful classifier from a series of weak classifiers. Boosting algorithms are critical in dealing with the bias-variance trade-off. Boosting algorithms, as opposed to packing algorithms, that only control for high variance in a model, control both aspects (bias and variance) and are thought to be more effective.

www.ijprems.com
editor@ijprems.com

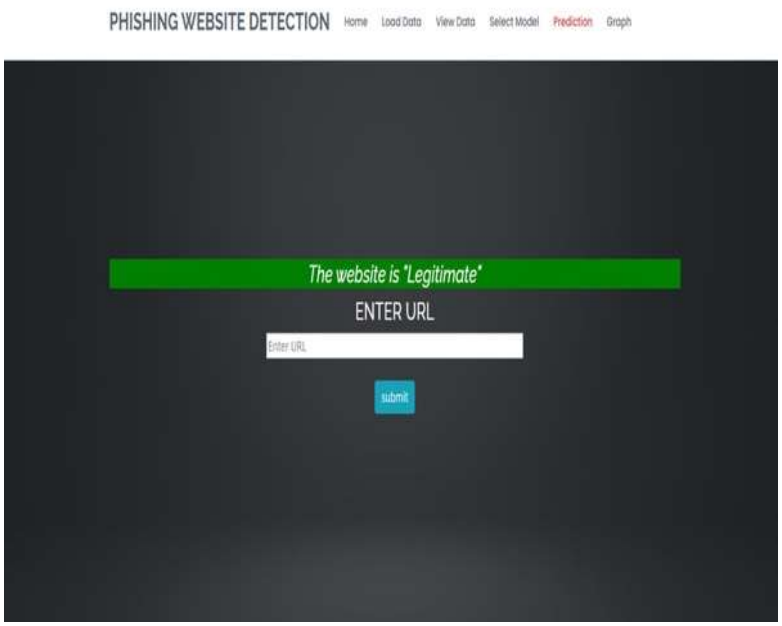
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 04, April 2023, pp : 515-518

e-ISSN : 2583-1062

Impact Factor : 5.725

5. EXPERIMENTAL RESULTS



6. FLOWCHART

7. ADVANTAGES

- It's more precise than the decision tree algorithm.
- It provides a useful strategy for handling missing data.
- It can produce a reliable prediction even without hyper-parameter tuning.
- It deals with the issue of overfitting in decision trees.
- The splitting point of each random forest tree chooses a subset of features.

8. CONCLUSION AND FUTURE SCOPE

This research provided multiple ML researchers' algorithms and techniques for spotting phishing websites. We came to the conclusion after reading the papers that the majority of the effort was accomplished using well-known algorithms like XGBoost, Decision Tree, and MLP classifier, which produces the neural network results. A new system for detection, like Phish Score and Phish Checker, was suggested by some authors. Accuracy, precision, recall, and other feature combinations were utilised. Certain characteristics may be added or outdated ones supplanted with new ones in order to detect the increasing number of phishing websites.

There are a lot of things that could be improved or added in the upcoming work.In this project, the ID3 and Naive Bayes classifiers are the two data mining classifiers we have chosen to use. There are additional classifiers, including the Bayesian network, neural network, and C4.5 classifiers. Such classifiers could be counted in the future to provide more data to be compared with, but they were not included in this study.

9. REFERENCES

[1] J. Shad and S. Sharma, "A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology," pp. 425–430, 2018.

[2] Y. Sönmez, T. Tuncer, H. Gökal, and E. Avci, "Phishing web sites features classification based on extreme learning machine," 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.

[3] T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301, 2018.

[4] M. Karabatak and T. Mustafa, "Performance comparison of classifiers on reduced phishing website dataset," 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.

[5] S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, "A New Method for Detection of Phishing Websites: URL Detection," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, vol. 0, no. Icicct, pp. 949–952.

[6] K. Shima et al., "Classification of URL bitstreams using bag of bytes," in 2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), 2018, vol. 91, pp. 1–5.

[7] A. Vazhayil, R. Vinayakumar, and K. Soman, "Comparative Study of the Detection of Malicious URLs Using Shallow and Deep Networks," in 2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018, 2018, pp.1– 6.

[8] W. Fadheel, M. Abusharkh, and I. Abdel-Qader, "On Feature Selection for the Prediction of Phishing Websites," 2017 IEEE 15th Intl Conf Dependable, Auton. Secur. Comput. 15th Intl Conf Pervasive Intell. Comput. 3rd Intl Conf Big Data Intell. Comput. Cyber Sci. Technol. Congr., pp. 871–876, 2017.

[9] X. Zhang, Y. Zeng, X. Jin, Z. Yan, and G. Geng, "Boosting the Phishing Detection Performance by Semantic Analysis," 2017.