Phishing URL Detection using Machine Learning

Abstract—**Most of the economic fraudulent falls under phishing attacks. it is one of the most vulnerable threats to online services and records. The significant factor is to improve one's common sense and a great deal of warning irrespective of any message that the user gets which seems even faintly suspicious and has tell-tale signs like orthographical errors or odd phrasing, mistakes done by malware authors regularly, urges you to do something immediately, or has a link or attachment which appears even remotely dodgy. A message from a trusted organization or user or even familiar faces could be spoofed. The proposed model is based on machine learning to know the classification of URL are extracted by means of its characteristics and behaviour. XGBoost is the model that uses an ensemble learning technique to gain knowledge on the data, a supervised algorithm which makes use of boosted approximations to obtain the best fit of the tree that uses the training dataset to the expectation of the target variable.**

Keywords— Phishing, Machine Learning, XG Boost, dataset, supervised and feature extraction.

I. INTRODUCTION

The world is evolving with cutting edge technology, the people around the world are interconnected with each other through the internet with the help of electronic devices and smart gadgets. There are about 5 billion active internet users worldwide. A major threat to the Internet users is cyberattacks. Phishing is the most common attack where the attackers use illegitimate websites to obtain victims data and use it illegally. Phishing can easily be imposed using the email and by other modes of communication. The attackers grab data easily and get valuable data. The phishing attacks has a huge impact on data loss, unknown malware injections, identity thefts and much more. This process is intended to get data such as user ID and user passwords, One Time Passwords through mail or text messages, payment details, and sensitive information.

The attackers attempt to gain the users trust by using the same user interface, almost same URL, and cloned websites. An efficient way to detect and prevent phishing is by using an automated approach. Machine learning helps to achieve this. It uses algorithms, statistical models, patterns, and inferences to complete certain tasks without human intervention. Machine learning enables learn and improve from its experience without being overtly programmed. The algorithm learns with the training datasets. The major advantage of the algorithm is to allow the system to learn and decide automatically whether the website is phishing or legitimate.

A phishing URL differs by various features from the original one. Sometimes, a small change such as spelling errors, extra characters and a missing character can result in a lookalike kind of replicas. It can be identified using a good common-sense and a basic knowledge in the actual websites. The domain names must be unique so that small changes can factor so much. The URL is segregated by its features to several categories to predict the nature of the URL.

II. PROPOSED METHODOLOGY

A. METHODOLOGY

A machine learning approach predicts the URL whether it is a phishing or not. It is achieved by the extracted features from the input URL. The features are classified by most common ways of identifying phishing links. The purpose of the proposed methodology is to obtain a high accuracy in detecting phishing sites over the Internet.

The quality of the prediction of a ML algorithm is strongly related to the quality of its training dataset. The Machine Learning approach requires a supervised learning algorithm, and therefore the samples need to be labelled as either benign or malicious. Firstly, the raw data of phishing and legitimate URLs are extracted from the official websites of “phishtank” and “openphish”, “pagerank” and “alexa”.

B. FEATURE EXTRACTION

The URLs are classified by its features. They are segregated by its features which consists of 30 characteristics. The features are based on the characteristics of phishing and legitimate URLs.

a) URL based features:

i. URL with respect to IP address: The URL must be active and take to an active IP address.

ii. Long URL to hide suspicious part: It has been a common observance that phishing web pages usually have long URLs that attempt to hide malicious URL fragments from the user. We take the assumption that a web page with a long URL is necessarily a phishing or suspicious site. In the event the assertion fails, i.e., for a legitimate web page with valid long URLs, the absence of other phishing attributes on the web page will balance the wrong assumption and correctly classify a legitimate web page as non-phishing.

iii. Use of URL shortening services: A shortened URL hides the real URL behind a redirection hop. A web page that uses a URL shortening service such as Tiny URL is highly suspicious and is likely to be a phishing attempt. Therefore, we set the rule that if the URL has been shortened using a URL shortening service, then it is a phishing page and legitimate otherwise.

iv. Use of "@" symbol: Needs verification The "@" symbol is a reserved keyword according to Web standards. So, the presence of "@" in a URL is suspicious and the web page is taken as phishing and legitimate otherwise.

v. Redirection with double shahs “//”: The presence of "//" in the address bar indicates that it will be redirected to another IP. If the double slash in the address bar is greater than seven, then might be a phishing site. Fixed number of seven repetitions may vary for every site and an average of seven is taken.

vi. The hyphen “- “: Usually, the phishing sites contains hyphens to split the words in an actual website to look like the reputed one.

vii. Sub domains and multi sub domains: If a URL has more than three dots in the domain part then it is considered as a phishing site and legitimate otherwise.

b) Abnormal based features:

viii. Request URL: A legitimate site usually has external page objects such as images, animations, files, etc. be accessed by a request URL which shares the same domain as the web page URL. We classify sites which fail this rule as phishing.

ix. URL portion of anchor tag: We check if the domain in the URL portion of all anchor tags match the main URL of the page and if the anchor tag has only URL fragments or JavaScript functions.

x. Links in <meta>, <script> and <link> tags: We check if the domain of the links in the <meta>, <script> and <link> tags match the domain in the mail URL.

xi. Server Form Handler (SFH): When a form is submitted, some valid action must be taken. So, if the action handler of a form is empty or "about: blank" or if the domain of the action URL is different from the domain of the main URL, then it is taken as a phishing site.

xii. Submitting Information to Email: If the webpage contains a "mailto:" function then it is taken as a phishing site and legitimate otherwise.

c) HTML and JavaScript based features:

xiii. Status bar customization: Phishers can modify the status bar using JavaScript to show a legitimate URL. By analysing the "onMouseOver" events in the web page we can determine if such a modification has occurred.

xiv. Disabling right click option: Phishers can disable the right click option to prevent the user from checking the source code of the page. This is verified by analysing the source code.

xv. Using pop-up window: Legitimate sites rarely ask for user info on a pop-up window, whereas phishing sites generally use pop-up windows to get user info.

xvi. “Iframe” redirection: Phishers also use “Iframe” tags with invisible borders to get user info and redirect to the original site. We analyse the source code to check if “Iframe” tags are used.

d) Domain based features:

There are about 14 domain-based features such as “alexa” “pagerank”, google page index, different “whois” results, etc. All the 30 segments of the features extracted from the address bar contains either of the three values: “0”, “1”, “-1”. Each value is the output from a processed program which determines it is a phishing or legitimate. The machine learning model decides how to process each and provides proper prediction.

III. MODEL SELECTION

The machine learning model has a specific type of processing methods. Every classifier has its own working algorithms. As the desired output is a prediction of type “this” or “that” type, a proper machine learning model is to be selected for the required output. A set of ML classifiers are taken to test the dataset. The machine learning algorithms used are Decision Tree, Random Forest, K Nearest Neighbours Classifier, Extreme Gradient Boost, and Support Vector Machine.

The models predicted the values which says phishing or not. The predicted values in each model’s accuracy are compared to select the high-performance model for our requirements. The testing data and evaluating the prediction with different machine learning algorithms and with the extended number of features gives more accuracy which is a bit more than other existing systems.

TABLE 1 – Accuracy Scores of models

ML Model	Train Accuracy	Test Accuracy
Decision Tree	0.993	0.919
Random Forest	0.935	0.919
KNN	0.999	0.882
XG Boost	0.997	0.940
SVM	0.927	0.922

a. The results of various machine learning models

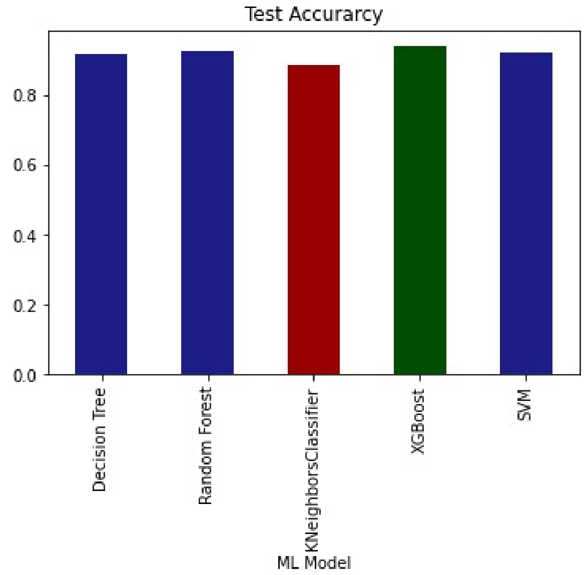

Fig.1. Test accuracy of the selected models

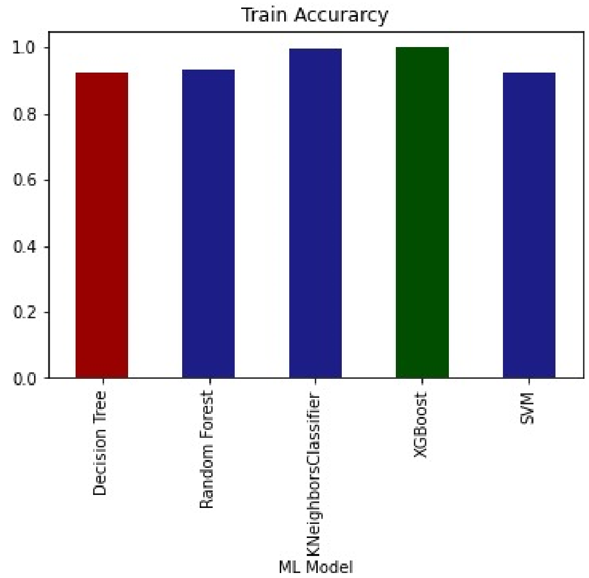

Fig.2. Train accuracy of the selected models

IV. RESULTS

The test results were obtained from the different machine learning models with the same dataset. The graphs of the results are given below. The URL is given as input to the system, it undergoes the feature extraction process specified in the Feature Extraction module. The extracted features will then go into the machine learning algorithm and predicts the input URL is a phishing or legitimate.

The algorithm which predicts with the most accuracy is the one which we need. To find the best performing algorithm, we need the results of different algorithms. Let's see the performance of each algorithm used in the project. As this is a supervised machine learning model, the predicted values are already known so that we can find the exact accuracy of the models. The Decision Tree algorithm shows train accuracy of 99.3% and test accuracy of 91.9%. The Random Forest algorithm which has the train accuracy of 93.5% and test accuracy of 91.9%. The "K-Nearest Neighbor" algorithm has the train accuracy of 99.9% and test accuracy of 88.2% while the Extreme Gradient Boosting algorithm has the train accuracy of 99.7% and test accuracy of 94.0% and The Support Vector Machine algorithm which has the train accuracy of 92.7% and test accuracy of 92.2%.

REFERENCES

[1] Joseph Johnson (2022). "Worldwide digital population as of April 2022" Published by, May 9, 2022.

[2] Gunter Ollmann (2004). "Securing against the 'threat' of instant" Published by Network Security Volume 2004, Issue 3, March 2004, Pages 8-11.

[3] Times of India (2021). "India among top 3 Asian nations affected by DNS cyber-attacks: Report". Updated: Jun 7, 2021, 17:04 IST.

[4] Suzanne Widup, Alex Pinto, Gabriel Bassett, David Hylender (2021) Verizon Data Breach Investigations Report, May 2021.

[5] Symbol Security (author) (2019). "Verizon Data Breach Investigations Report 2019". May 30, 2019.

[6] WidupSuzanne, WidupMarc, SpitlerDavid Hylender Gabriel (2018). "Verizon Data Breach Investigations Report 2018". April 2018.

[7] Moubayed A, Mohammadnoor AM Injadat, Mohammadnoor AM Injadat, Ali B Nassif (2018). "Challenges and Research Opportunities Using Machine Learning & Data Analytics", July 2018 IEEE Access PP (99):1-1.

[8] Kintis P, Miramirkhani N, Lever C, Chen Y, Romero-Gomez, R, Pitropakis, N, Nikiforakis, N and Antonakakis, M (2017). "Hiding in Plain Sight: A Longitudinal Study of Combosquatting Abuse. Association of Computer Machinery's", Computer and Communications Security (ACM CCS) Dallas, Texas USA 30 Oct - 02 Nov 2017.

[9] Sagar P, Yogesh S, Nilesh S (2020). "Detecting Phishing Websites Using Machine Learning". IRJET e-ISSN: 2395-0056 Volume: 07 Issue: 02, Feb 2020.

[10] Dr. SD Sawarkar, Sophiya Shikalgar, Mrs. Swati Narwane (2019). "Detection of URL based Phishing Attacks using Machine Learning ". IJERT ISSN: 2278-0181 Vol-8 Issue-11, November 2019.