Machine learning approaches for phishing websites detection

Submitted To:
Dr. Ankit Kumar Jain
Assistant Professor
National Institute of Technology
kurukshetra(Haryana)- 136119

Submitted By:
MD SAIF ALI
32013112
M.TECH(COMPUTERENGINEERING)

Introduction

Phishing is an online identity theft, which can deceive Internet users into revealing their secret information and credentials, e.g., login id, password, credit card number, etc

Phishing is one of the major computer security threats faced by the cyber-world and could lead to financial losses for both industries and individuals.

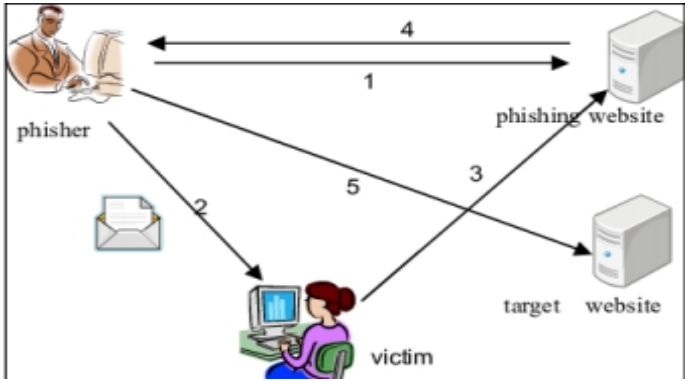


Fig-Steps of Web phishing process.

Objective of the Project

phishing is the fraudulent attempt to obtain sensitive information such as username, password, bank account details, and credit card details for malicious use. Phishing frauds might be the most popular cybercrime used today. There are variousdomains where phishing attack can occur like online payment sector, webmail, and financial institution, file hosting or cloud storage and many others. We have developed a system that uses machine learning techniques to classify websites based on their URL. The objective of this project is to train machine learning models and deep neural networks on the dataset created to predict phishing websites. The models were tested with a data set containingthousands of real world URLs where each could be categorized as a legitimate site,suspicious site, or phishing site

Overview of phishing detection approaches proposed in the literature

Phishing detection approaches split into two classes:

1. User education based approaches

User education approach aims to improve the capacity of Internet users in the detection of phishing attacks Internet users can be educated to distinguish the characteristics of phishing and legitimate emails and websites.

2. Software-based approaches

1. Phishing blacklist: A blacklist contains the list of malicious domains, URLs and IP addresses. The blacklist needs to be the regular update from their source because thousands of fake websites launch every day

2.Visual similarity based Techniques: These techniques utilize various features to compute the similarity between websites like page source code, images,textual content, text formatting, HTML tags, CSS, website logo, etc. Most of the visual similarity based approaches compare the new website with previously visited or storedwebsites. Therefore, these techniques cannot detect the new phishing websites and produce high false negative rate.

3. Search Engine based techniques: The search Engine (SE) based techniques extract identity features (e.g., title, copyright, logo, domain name, etc.) from the webpage and make use of the search engine to check the legitimacy of webpage.

4.Machine Learning based techniques: These methods train a machine learning algorithm with some features that can distinguish a genuine website from the phishing one. In this, a website is declared as phishing, if the design of the websites matches with the predefined feature set. The performance of these solutions depends on features set, training data and classification algorithm. These features are extracted from various sources like URL, page source, website traffic, search engine, DNS, etc

MACHINE LEARNING MODELS

1. Decision Tree: Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

2.Random Forest: Random Forest (RF) is another popular decision tree, which can be used for both classification and regression. RF is an ensemble of a number of decision trees independently trained on selected training datasets. The classification information is then determined by voting among all the trained decision trees. Therefore, Random Forest usually achieves a better classification accuracy compared to a single tree.

MACHINE LEARNING MODELS(Contd.)

3.Multilayer Perceptrons: Multilayer perceptrons (MLPs) are also known as (vanilla) feed-forward neural networks, or sometimes just neural networks. Multilayer perceptrons can be applied for both classification and regression problems. MLPs can be viewed as generalizations of linear models that perform multiple stages of processing to come to a decision.

4.Support Vector Machines: The support vector machine (SVM) is one of the most well known and robust supervised machine learning techniques, which has been utilized effectively in many science and engineering applications. SVM is based on maximizing the margin and thereby creating the largest possible distance between the separating hyperplane and the instances to reduce an upper bound on the expected generalization error.

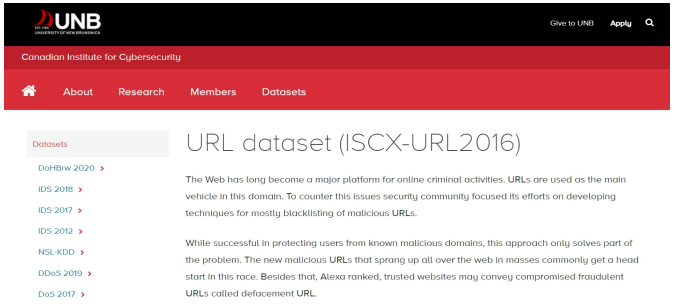
APPROACH

Below mentioned are the steps involved in the completion of this project:

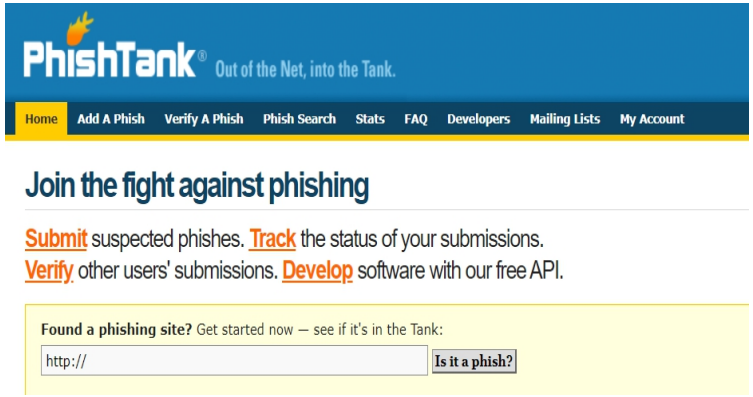
1. Collect dataset containing phishing and legitimate websites from the open source platforms.
2. Write a code to extract the required features from the URL database.
3. Analyze and preprocess the dataset by using EDA techniques.
4. Divide the dataset into training and testing sets.
5. Run selected machine learning algorithms like Decision tree, SVM, Random Forest,etc
6. Write a code for displaying the evaluation result considering accuracy metrics.
7. Compare the obtained results for trained models and specify which is better.

DATA COLLECTION

- Legitimate URLs are collected from the dataset provided by University of New Brunswick, https://www.unb.ca/cic/datasets/url-2016.html
  - From the collection, 10,000 URLs are randomly picked.



- Phishing URLs are collected from opensource service called PhishTank .This service provide a set of phishing URLs in multiple formats like csv, json etc. that gets updated hourly. https://www.phishtank.com
  - Form the obtained collection, 10,000 URLs are randomly picked



FEATURE SELECTION

The following category of features are selected:

1.URL based features
2.Login form based features
3.Hyperlink specific based features
4.CSS based features

1.URL based features

1. Number of dots(.) in URL (https://support.appleid.itune.com-txutwo3wfh.store)
2. Presence of Special symbol in URL.(www.paypal-india.com)
3. Lenght of URL.
4. Suspicious word in URL
5. Position of Top-Level domain(http://support.paypal.com.prodigitalmedia.org/signin/?country.x=US\&loc,)
6. http count in URL
7. Brand name in URL (http://forlittledrops.org/asd/Paypalaccount/)

FEATURE SELECTION (Contd.)

2.Login form based features

1.Fake login form.

3.Hyperlink specific based features

1. Number of webpages
2. No hyperlink feature
3. Foreign hyperlinks
4. Empty hyperlinks
5. Error in hyperlinks
6. Hyperlinks redirection

4.CSS based features

1. Copied CSS(Cascading Style Sheets)

MODEL EVALUATION

The models are evaluated, and the considered metric is accuracy.

Below Figure shows the training and test dataset accuracy by the respective models:

	ML Model	Train Accuracy	Test Accuracy
2	Multilayer Perceptrons	0.864	0.850
1	Random Forest	0.821	0.814
3	SVM	0.802	0.802
0	Decision Tree	0.814	0.801

- For the above it is clear that the Multilayer Perceptrons model gives better performance

Conclusion

Phishers use spoofed e-mail, phishing software to steal personal information and financial account details such as usernames and passwords. This project is mainly deals with methods for detecting phishing websites by analyzing various features of benign and phishing URLs by using Machine learning techniques.

REFERENCES

[1]. A. K. Jain and B. B. Gupta, “Comparative analysis of features based machine learning approaches for phishing detection”
https://ieeexplore.ieee.org/abstract/document/7724641

[2]. AK Jain, BB Gupta, “Towards detection of phishing websites on client-side using machine learning based approach.”
https://link.springer.com/article/10.1007/s11235-017-0414-0

[3]. Waleed Ali, “Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection”

THANK YOU