Project Design Phase-I

Proposed Solution Template

Date	22 October 2022
Team ID	PNT2022TMID27211
Project Name	WEB PHISING DETECTION
Maximum Marks	2 Marks

Proposed Solution Template:

Project team shall fill the following information in proposed solution template.

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	Cyber criminals use phishing emails because it's easy, cheap and effective. Email addresses are easy to obtain, and emails are virtually free to send. With little effort and cost, attackers can quickly gain access to valuable data. Attachments from phishing emails can contain malware that once opened can leave the door open to the attacker to perform malicious behavior from the user's computer.
2.	Idea / Solution description	Even though there are several methods exists today to detect phishing but still it has become a very difficult task to detect fake E-mails in the current scenario. Today there are a number of techniques exist for identification of phishing E-mails and some of them are white listing, heuristics, blacklisting and machine learning. A machine learning technique is proposed in this chapter to identify the phishing E-mails and protect the user from revealing their pin, user id and passwords. The objective of this chapter is to use J48 one of the machine learning algorithms to analyze incoming E-mails and helps in preventing the user from phishing attacks. This chapter presented an architectural model as shown in Figure 1 below and uses the various sub-processes at different stages to classify between fake E-mail and genuine E-mails.

3.	Novelty / Uniqueness	Phishing is a form of fraud in which the attacker tries to learn sensitive information such as login credentials or account information by sending as a reputable entity or person in email or other communication channels. Typically a victim receives a message that appears to have been sent by a known contact or organization. The message contains malicious software targeting the user's computer or has links to direct victims to malicious websites in order to trick them into divulging personal and financial information, such as passwords, account IDs or credit card
4.	Social Impact / Customer Satisfaction	There are a lot of algorithms and a wide variety of data types for phishing detection in the academic literature and commercial products. A phishing URL and the corresponding page have several features which can be differentiated from a malicious URL. For example; an attacker can register long and confusing domain to hide the actual domain name **(**Cybersquatting, Typosquatting). In some cases attackers can use direct IP addresses instead of using the domain

		This type of event is out of our scope, but it can be used for the same purpose. Attackers can also use short domain names which are irrelevant to legitimate brand names and don't have any FreeUrl addition. But these type of web sites are also out of our scope, because they are more relevant to fraudulent domains instead of phishing domains. Beside URL-Based Features, different kinds of features which are used in machine learning algorithms in the detection process of academic studies are used
5.	Business Model (Revenue Model)	There are a lot of algorithms and a wide variety of data types for phishing detection in the academic literature and commercial products. A phishing URL and the corresponding page have several features which can be differentiated from a malicious URL. For example; an attacker can register long and confusing domain to hide the actual domain name **(**Cybersquatting, Typosquatting). In some cases attackers can use direct IP addresses instead of using the domain name.

		This type of event is out of our scope, but it can be used for the same purpose. Attackers can also use short domain names which are irrelevant to legitimate brand names and don't have any FreeUrl addition. But these type of web sites are also out of our scope, because they are more relevant to fraudulent domains instead of phishing domains. Beside URL-Based Features, different kinds of features which are used in machine learning algorithms in the detection process of academic studies are used. Features collected from academic studies for the phishing domain detection with machine learning techniques are grouped as given below.
6.	Scalability of the Solution	Collecting legitimate domains is another problem. For this purpose, site reputation services are commonly used. These services analyse and rank available websites. The websites which have high rank scores are legitimate sites which are used very frequently. When we have raw data for phishing and legitimate sites, the next step should beInitially, as we mentioned above, phishing domain is one of the classification problem. When we calculate the features that we've selected our needs and purposes, our dataset looks like in figure below

In our examples, we selected 12 features, and we calculated them. Thus we generated a dataset which will be used in training phase of machine learning algorithm.

No.	1: domain String	2: tld String	3: brandName Numeric	4: editDbrandName Numeric	5: digitCount Numeric	6: length Numeric	7: isKnownTld Numeric	8: www Numeric	9: keywords Numeric	10: punnyCode Numeric	11: randomDomain Numeric	12:
...	ayanasalon	com	0.0	1.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	
...	esteticabrasilbeauty	com	0.0	1.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	
...	erate365	com	0.0	1.0	3.0	8.0	0.0	0.0	0.0	0.0	0.0	
...	upstatescbusiness	com	1.0	1.0	0.0	17.0	0.0	0.0	1.0	0.0	0.0	
...	6--4c	com	0.0	0.0	2.0	4.0	0.0	0.0	0.0	0.0	0.0	
...	services--confirmatio...	com	1.0	1.0	0.0	23.0	0.0	0.0	1.0	0.0	0.0	
...	hmsinformatica	com	1.0	1.0	0.0	14.0	0.0	0.0	1.0	0.0	0.0	

A Decision Tree can be considered as an improved nested-if-else structure. Each features will be checked one by one. An example tree model is given below. Generating a tree is the main structure of detection mechanism. Yellow and elliptical shaped ones represent features and these are called nodes. Green and angular ones represent classes and these are called leaves. The *length* is checked when an example arrives and then the other features are checked according to the result. When the journey of the samples is completed, the class that a sample belongs to will become clear.