

Project report

Project Name:

Web Phishing Detection

Team Id:

PNT2022TMID47961

Team Members:

S.Negavarshini(Team Leader)

M.Alagulakshimi

R.Durgadevi

K.Gayathri

INDEX

1. INTRODUCTION

- 1.1 Project Overview
- 1.2 Purpose

2. LITERATURE SURVEY

- 2.1 Existing problem
- 2.2 References
- 2.3 Problem Statement Definition

3. IDEATION & PROPOSED SOLUTION

- 3.1 Empathy Map Canvas
- 3.2 Ideation & Brainstorming
- 3.3 Proposed Solution
- 3.4 Problem Solution fit

4. REQUIREMENT ANALYSIS

- 4.1 Functional requirement
- 4.2 Non Functional requirements

5. PROJECT DESIGN

- 5.1 Data Flow Diagrams
- 5.2 Solution & Technical Architecture

6. PROJECT PLANNING & SCHEDULING

- 6.1 Sprint Planning& Estimation
- 6.2 Sprint delivery schedule

7. CODING & SOLUTIONING (Explain the features added in the project along with code)

- 7.1 Coding
- 7.2 solution

8. TESTING

9. RESULTS

10. ADVANTAGES & DISADVANTAGES

11. CONCLUSION

12. FUTURE SCOPE

13. APPENDIX

Source Code

GitHub & Project Demo Link

1.1. INTRODUCTION

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website.

Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. In United States businesses, there is a loss of US\$2billion per year because their clients become victim to phishing [1]. In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as \$5 billion [2]. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques.

Overview:

A phishing website is a common social engineering method that mimics trustful uniform resource locators (URLs) and webpages. The objective of this project is to train machine learning models on the dataset given to predict phishing websites. Both phishing and benign URLs of websites are gathered to form a dataset and from them required URL and website content-based features are extracted. The performance level of each model is measured and compared.

Purpose:

A phishing campaign is an email scam designed to steal personal information from victims. Cybercriminals use phishing, the fraudulent attempt to obtain sensitive information such as credit card details and login credentials, by disguising as a trustworthy organization or reputable person in an email communication.

LITERATURE SURVEY:

The purpose or goal behind phishing is data, money or personal information stealing

through the fake website. The best strategy for avoiding the contact with the phishing web

site is to detect real time malicious URL. Phishing websites can be determined on the basis

of their domains. They usually are related to URL which needs to be registered (low-level

domain and upper-level domain, path, query). Recently acquired status of intra-URL

relationship is used to evaluate it using distinctive properties extracted from words that

compose a URL based on query data from various search engines such as Google and Yahoo.

These properties are further led to the machine-learningbased classification for the

identification of phishing URLs from a real dataset. This paper focus on real time URL

phishing against phishing content by using phish-STORM. For this a few relationship

between the register domain rest of the URL are consider also intra URL relentless is

consider which help to dusting wish between phishing or non phishing URL. For detecting a

phishing website certain typical blacklisted urls are used, but this technique is unproductive

as the duration of phishing websites is very short. Phishing is the name of avenue. It can be

defined as the manner of deception of an organization's customer to communicate with

their confidential information in an unacceptable behaviour. It can also be defined as

intentionally using harsh weapons such as Spasm to automatically target the victims and

targeting their private information. As many of the failures being occurred in the SMTP are

exploiting vectors for the phishing websites, there is a greater availability of communication

for malicious message deliveries.

Proposed a novel classification approach that use heuristic based feature extraction

approach.

In this, they have classified extracted features into different categories such as URL

Obfuscation features, Hyperlink-based features.

Moreover, proposed technique gives 92.5% accuracy. Also this model is purely depends on

the quality and quantity of the training set and Broken links feature extraction.

REFERENCES

- Liu J, Ye Y (2001) Introduction to E-business operators: commercial center arrangements, security issues, and market interest. In: E-business specialists, commercial center arrangements, security issues, and market interest, London, UK

- APWG, Aaron G, Manning R (2013) APWG phishing reports. APWG, 1 February 2013.

[Online]. Accessible: [http://www. antiphishing.org/assets/apwg-reports/](http://www.antiphishing.org/assets/apwg-reports/).
Gotten to 8

Feb2013

- Kaspersky Lab (2013) Spam in January 2012: love, governmental issues and game.

[Online].

Available:<http://www.kaspersky.com/about/news/spam/2012>

Spam_in_January_2012_Love_Politics_and_Sport. Gotten to 11 Feb 2013

- Seogod (2011) Black Hat SEO. Search engine optimization Tools. [Online].

Accessible:[http://www.seobesttools.com/dark cap website optimization/](http://www.seobesttools.com/dark%20cap%20website%20optimization/). Gotten to 8 Jan

2013

- Dhamija R, Tygar JD, Hearst M (2006) Why phishing works. In: Proceedings of the SIGCHI meeting on human factors in figuring frameworks, Cosmopolitan Montre 'al, Canada

- Cranor LF (2008) A system for thinking about the human tuned in. In: UPSEC'08 Proceedings of the first meeting on ease of use, brain science, and security, Berkeley, CA,USA

- Miyamoto D, Hazeyama H, Kadobayashi Y (2008) An assessment of AI based techniquesfor recognition of phishing destinations. Aust J Intell Inf Process Syst 10(2):54–6

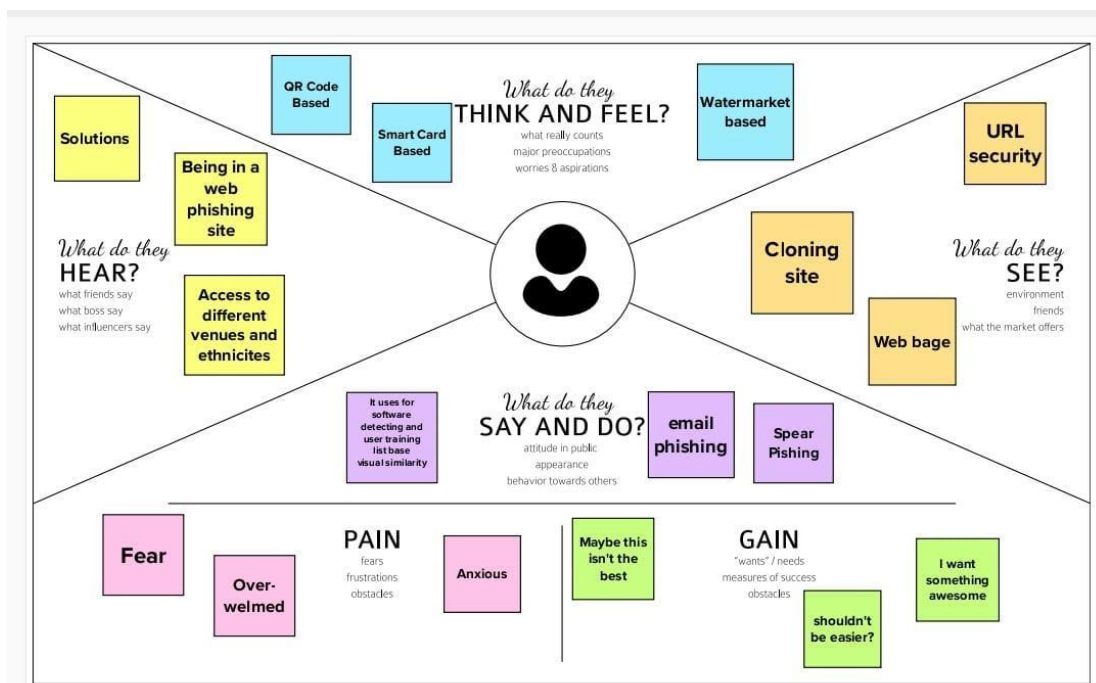
- Xiang G, Hong J, Rose CP, Cranor L (2011) CANTINA?: a include rich AI structure

for identifying phishing sites. ACM Trans Inf Syst Secur 14(2):1–28

Problem statement:

Phishing detection techniques do suffer low detection accuracy and high false alarm especially when novel phishing approaches are introduced. Besides, the most common technique used, blacklist-based method is inefficient in responding to emanating phishing attacks since registering new domain has become easier, no comprehensive blacklist can ensure a perfect up-to-date database. Furthermore, page content inspection has been used by some strategies to overcome the false negative problems and complement the vulnerabilities of the stale lists. Moreover, page content inspection algorithms each have different approach to phishing website detection with varying degrees of accuracy.

3. Ideation and proposed solution:



Ideation and brain storming:

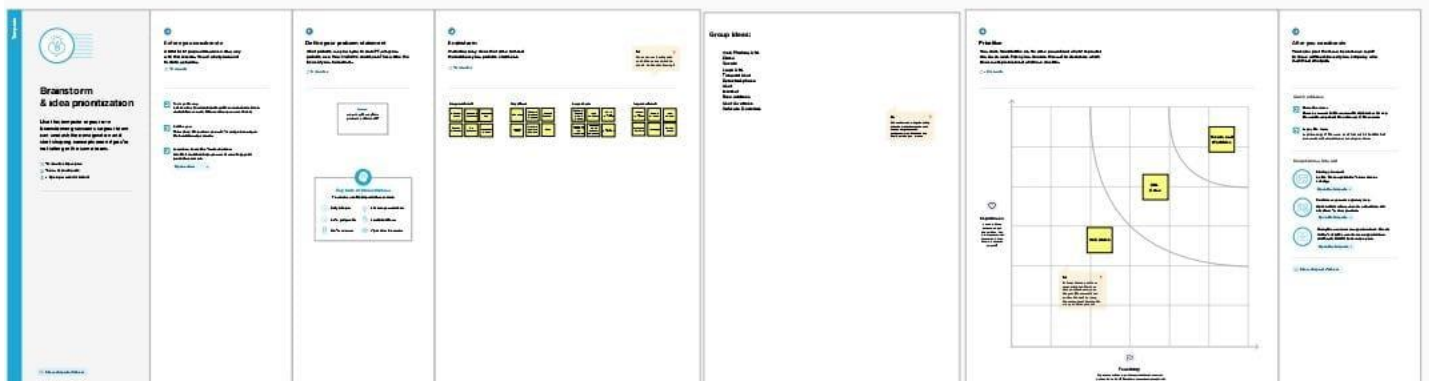
To minimize downtime and the impact to your brand, your security team needs a comprehensive strategy that encompasses security tools and processes across all the stages of an attack.

Register for the Gartner report and learn more to prepare your organization:

The attack pattern of a typical ransomware attack

Steps to defend and respond to a ransomware attack

Security tools and controls to consider



3.3 Proposed solution:

Date	19 September 2022
Team ID	PNT2022TMD47961
Project Name	Project – Web Phishing 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)	web phishing detection using machine learning a web URL suspected to be a phishing website.goal: To design an efficient and adaptive phishing detection algorithm that has the ability to adapt to newer data or phishing attack vectors discover in the wild
2.	Idea / Solution description	The data for this project is a collection of records. This stage includes choosing a sample of all available information on which to work. Data, especially as the huge quantity of data whereby the target output has been established, is the starting point for machine learning challenges. Data that has been labeled contains information for which we have an answer
3.	Novelty / Uniqueness	the uniqueness of Web SSO phishing, and thus have not been taken seriously by the community in the process of promoting the Web threat from phishing attacks to real-world Web SSO
4.	Social Impact / Customer Satisfaction	users' login credentials, financial information (such as credit cards or bank accounts), company data, and anything that could potentially be of value.
5.	Business Model (Revenue Model)	phishing detection model we suggest that all models should have FPR/FNR trade-off tuned to the particular risk assessment of the deployment site. In ... Hence in our project we focused
6.	Scalability of the Solution	This paper presents a proposal for scalable detection and isolation of phishing. The main ideas are to move the protection from end users towards the network provider and to employ the novel bad neighbourhood concept,

Problem solution fit:

Financial institutions are the most targeted of all industry sectors analyzed in the 2020 X-Force Threat Intelligence Index. IBM Security Trusteer® Rapport is an advanced endpoint protection solution designed to protect users from financial malware and phishing attacks.

Project Title: NEWS TRACKER APPLICATION		Project Design Phase-I - Solution Fit Template		Team ID: PNT2022TMID47959	
Define CS fit into CC	1. CUSTOMER SEGMENT(S) Who is your customer ? Our customers are who are interested to know about the social activities and who are all interested to know about the current trends in the society.	6. CUSTOMER CONSTRAINTS What constraints prevent your customers from taking action or limit their choices of solutions? I spending more time to gather the real information,Fake news,Non-relatable news,irrelevant informaton	5. AVAILABLE SOLUTIONS Which solutions are available to the customers when they face the problem or need to get the job done? What have they tried in the past? What pros & cons do these solutions	Explore AS, differentiate	
	2. JOBS-TO-BE-DONE / PROBLEMS Which jobs-to-be-done (or problems) do you address for your customers? There could be more than one; explore different sides. In current News tracking applications fake news spread fast and easily.Then,the wrong information confuse the people and they can't easily understand what the real information .	9. PROBLEM ROOT CAUSE What is the real reason that this problem exists? What is the backstory behind the need to do this job? Spreading the fake information, Too much of time will be taken for gather the information are real reason for this problem.	7. BEHAVIOUR What does your customer do to address the problem and get the job done? Find th correct application ,installation		
Focus on J&P, map into BE, understand RC				Focus on J&P, map into BE, understand RC	
- I d e e n t i f y t h e p r o b l e m					
3. TRIGGERS What triggers customers to act? Seeing their friends,colleagues,neighbours installing the application to know about the real Information.		10. YOUR SOLUTION To create a cloud application to get the information about the News.		8. CHANNELS of BEHAVIOUR 8.1 ONLINE What kind of actions do customers take online? Extract online channels from #7 Install the application	
4. EMOTIONS: BEFORE / AFTER How do customers feel when they face a problem or a job and afterwards? Confuse,not sure > confident,Sure,satisfaction					

Requirement analysis.

4.1 Functional requirements:

Solution Requirements (Functional & Non-functional)

Date	03 October 2022
Team ID	PNT2022TMID47961
Project Name	Project – Web phishing detection
Maximum Marks	4 Marks

Functional Requirements:

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Registration through Form Registration through Gmail Registration through LinkedIn
FR-2	User Confirmation	Confirmation via Email Confirmation via OTP
FR-3		
FR-4		

Non-functional Requirements:

Following are the non-functional requirements of the proposed solution.

FR No.	Non-Functional Requirement	Description
NFR-1	Usability	Share threat intelligence Remain alert about the latest threats and phishing attack tactics in your industry by sharing threat intelligence with partners and networks.
NFR-2	Security	Obtained centralized visibility to detect, investigate and respond to your most critical organization-wide cyber security threads with SIEM
NFR-3	Reliability	Our research demonstrates that current phishing detection technologies have an accuracy rate between 70% and 92.52%. The experimental results prove that the accuracy rate of our proposed model can yield up to 95%, which is higher than the current technologies for phishing website detection

NFR-4	Performance	Rules were extracted from the Random Forest Model and embedded into a Google chrome browser extension called PhishNet. PhishNet is built during the course of this research using web technologies such as HTML, CSS, and Javascript. As a result, PhishNet facilitates highly efficient phishing detection for the web.
NFR-5	Availability	In some cases, it may not be useful to use some of these, so there are some limitations for using these features. For example, it may not be logical to use some of the features such as Content-Based Features for the developing fast detection mechanism which is able to analyze the number of domains between 100.000 and 200.000. Another example would be, if we want to analyze new registered domains Page-Based Features is not very useful. Therefore, the features that will be used by the detection mechanism depends on the purpose of the detection mechanism. Which features to use in the detection mechanism should be selected carefully.
NFR-6	Scalability	This paper presents a proposal for scalable detection and isolation of phishing. The main ideas are to move the protection from end users towards the network provider and to employ the novel bad neighbourhood concept, in order to detect and isolate both phishing e-mail senders and phishing web servers.

5. Project design:

5.1 Data flow diagram:

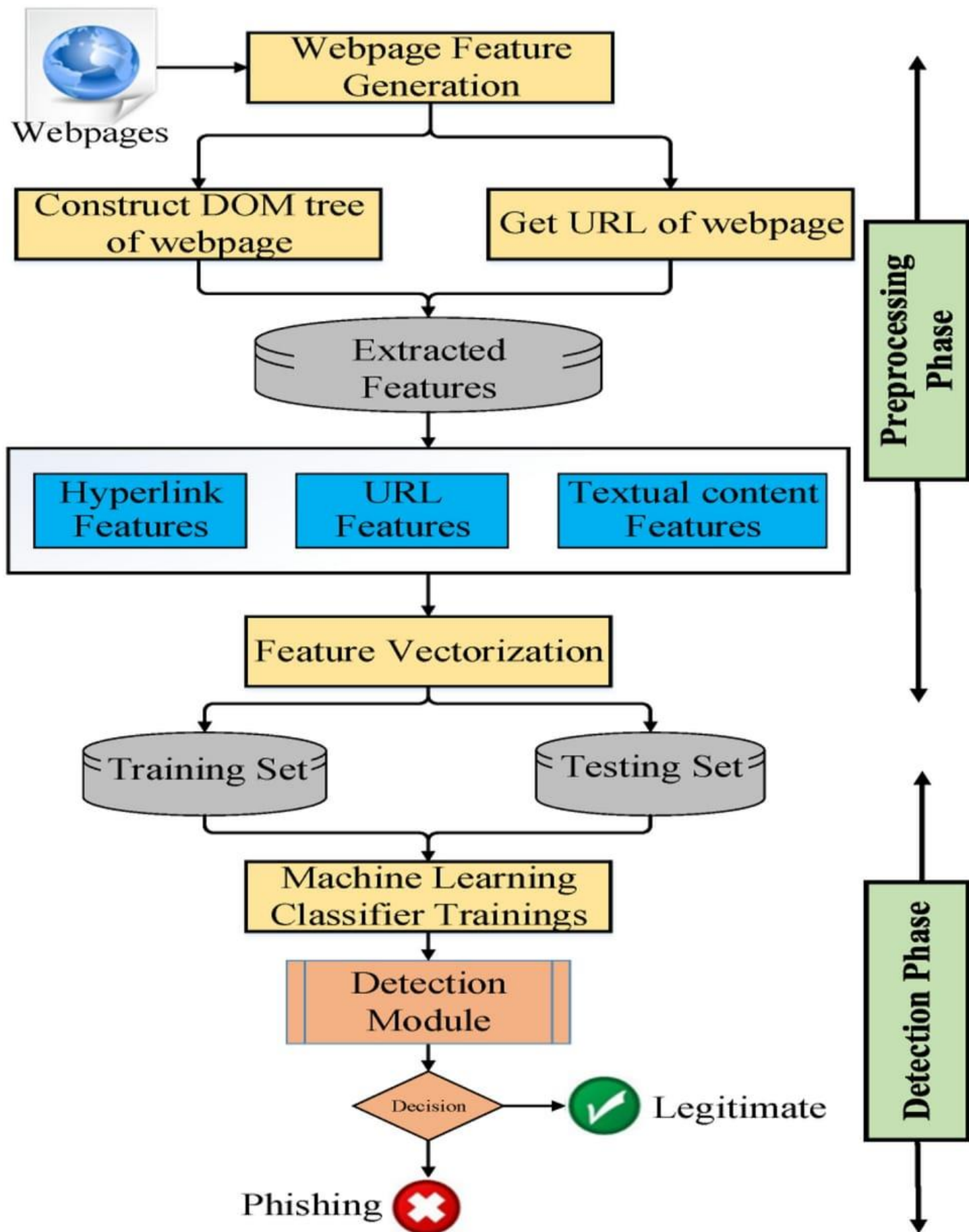


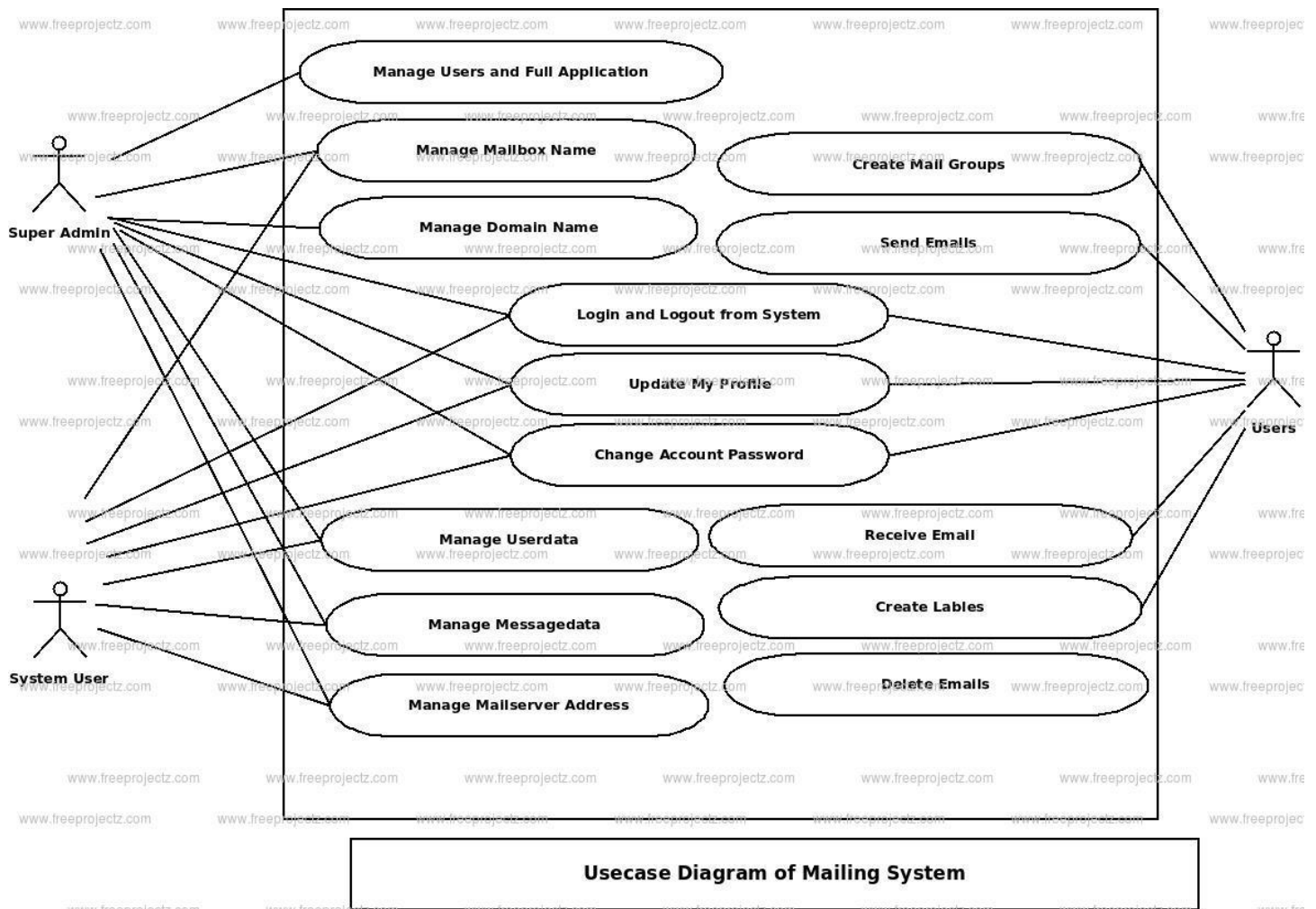
Table 1: components and technology:

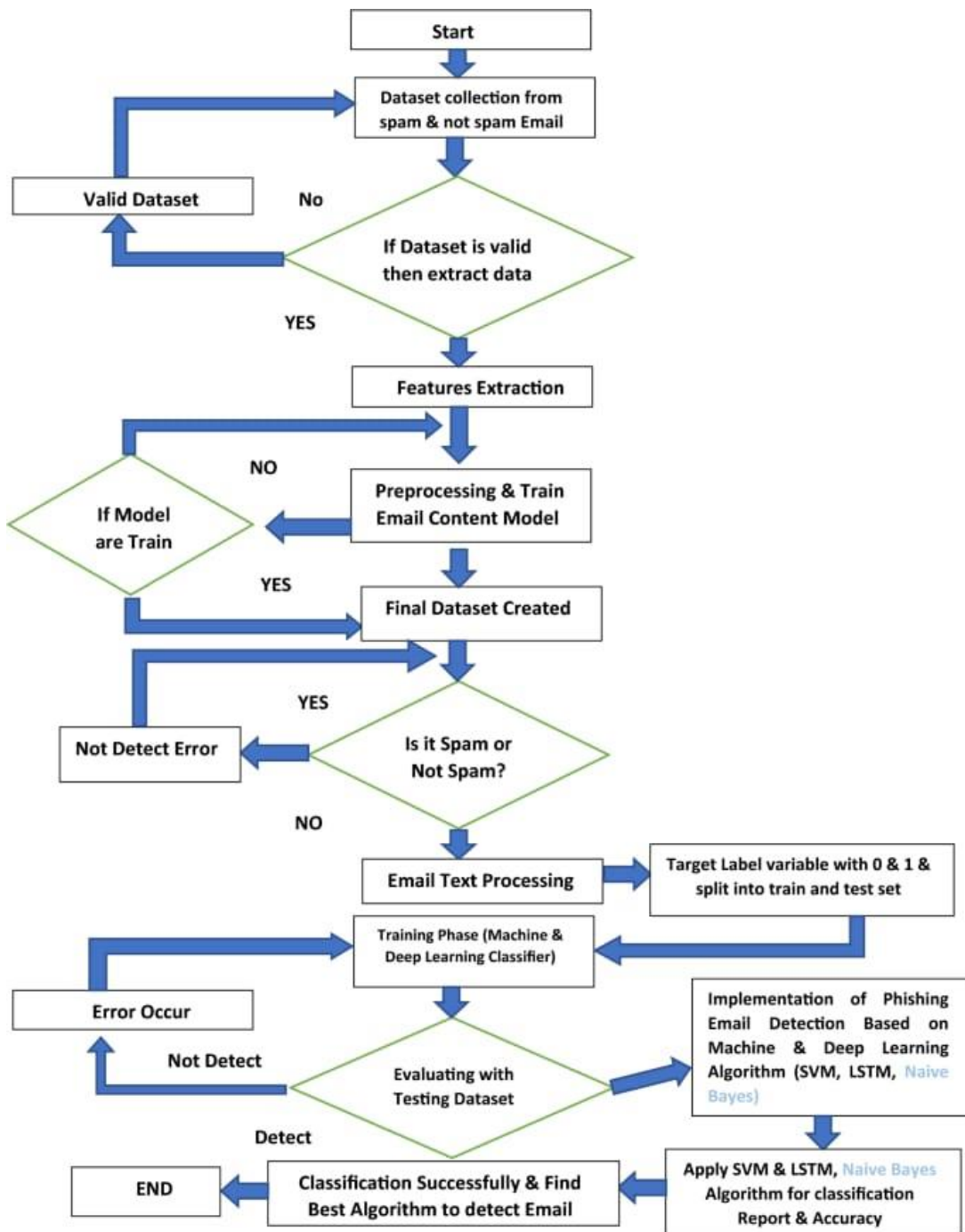
Sprint delivery plan

Date	18 November 2022
Team ID	PNT2022TMID47961
Project Name	SmartFarmer - Web phishing detection Application
Maximum Marks	4 Marks

Product Backlog, Sprint Schedule, and Estimation (4 Marks)

	Functional Requirement (Epic)	User Story Number		Points		Team Members
Sprint-1	Simulation creation	USN-1	Website application response times at different connection speeds Load test your web application to determine its behavior under normal and peak loads	2	High	Negavarshini, Alagulakshmi, Durgadevi, Gayathri.
Sprint-2	Software	USN-2	Creating device in the IBM Watson machine learning platform, workflow for machine learning	2	High	Negavarshini, Alagulakshmi, Durgadevi, Gayathri.





Solution and technical architecture:

Sprint-3	MIT App Inventor	USN-3	Test if a crash occurs due to peak load, how does the site recover from such an event	2	High	Negavarshini, Alagulakshmi, Durgadevi, Gayathri.

Sprint	User Story / Task		Story Priority			
Sprint-3	Dashboard	USN-3	Design the Modules and test the app	2	High	Negavarshini, Alagulakshmi, Durgadevi, Gayathri.
Sprint-4	analysis	USN-4	The analyst analyzes the email The analyst navigates to the web page of ThePhish and clicks on the "List emails" button to obtain the list of emails to analyze.	2	High	Negavarshini, Alagulakshmi, Durgadevi, Gayathri.

	Total Story Points	n	Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	7 Days	30 Oct 2022	06 Nov 2022	20	29 Oct 2022
Sprint-2	20	9 Days	31 Oct 2022	09 Nov 2022		05 Oct 2022

Project Tracker, Velocity & Burndown Chart: (4 Marks)

Sprint-3	20	6 Days	06 Nov 2022	13 Nov 2022		12 Oct 2022
Sprint-4	20	6 Days	11 Nov 2022	17 Nov 2022		15 Oct 2022

Start Velocity:

Imagine we have a 10-day sprint duration, and the velocity of the team is 20 (points per sprint). Let's calculate the team's average velocity (AV) per iteration unit (story points per day)

$$AV = \frac{\text{sprint duration}}{\text{velocity}} = \frac{20}{10} = 2$$

6. Project planning and scheduling:

Solution Requirements (Functional & Non-functional)

Date	03 October 2022
Team ID	PNT2022TMID47961
Project Name	Project – Web phishing detection
Maximum Marks	4 Marks

Functional Requirements:

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Registration through Form Registration through Gmail Registration through LinkedIn
FR-2	User Confirmation	Confirmation via Email Confirmation via OTP
FR-3		
FR-4		

Non-functional Requirements:

Following are the non-functional requirements of the proposed solution.

FR No.	Non-Functional Requirement	Description
NFR-1	Usability	Share threat intelligence Remain alert about the latest threats and phishing attack tactics in your industry by sharing threat intelligence with partners and networks.
NFR-2	Security	Obtained centralized visibility to detect, investigate and respond to your most critical organization-wide cyber security threats with SIEM
NFR-3	Reliability	Our research demonstrates that current phishing detection technologies have an accuracy rate between 70% and 92.52%. The experimental results prove that the accuracy rate of our proposed model can yield up to 95%, which is higher than the current technologies for phishing website detection

NFR-4	Performance	Rules were extracted from the Random Forest Model and embedded into a Google chrome browser extension called PhishNet. PhishNet is built during the course of this research using web technologies such as HTML, CSS, and Javascript. As a result, PhishNet facilitates highly efficient phishing detection for the web.
NFR-5	Availability	In some cases, it may not be useful to use some of these, so there are some limitations for using these features. For example, it may not be logical to use some of the features such as Content-Based Features for the developing fast detection mechanism which is able to analyze the number of domains between 100.000 and 200.000. Another example would be, if we want to analyze new registered domains Page-Based Features is not very useful. Therefore, the features that will be used by the detection mechanism depends on the purpose of the detection mechanism. Which features to use in the detection mechanism should be selected carefully.
NFR-6	Scalability	This paper presents a proposal for scalable detection and isolation of phishing. The main ideas are to move the protection from end users towards the network provider and to employ the novel bad neighbourhood concept, in order to detect and isolate both phishing e-mail senders and phishing web servers.

7. Coding and solution:

Phishing Url Detection Using PYTHON

Python:

```
import os;

from flask import flask

from flask import(
                    flash,render_template,request
)

from utils, import secure_filename

import phishing_detection

app=flask(__name__)

UPLOAD_FOLDER="/files"

app.config['UPLOAD_FOLDER']= UPLOAD_FOLDER

ALLOWED_EXTENSIONS=set(['txt','pdf','png'])

def allowed_file (filename) and /

filename.rsplit('.',1)

@app.route(/result)

def result();

    url name =request.args['name']

result =phishing_detecti.getResult(urlname)
```

```

return result

@app.route(¥result)

def result():
    urlname=request.args['name']
    result=phishing_detection.getResult(url name)
    return result

list=[X.start(0) for x in re.finditer()]
i=0 sucess=0
if soup==.find_all('link',href=True):
    dots=[x.start(0)for x in re.finditer
    if url in link['href'] or domain inlink[href']
    sucess=sucess+1
    i=i+1
    for script in soup.find all('script',src=true)
    def generate_data_set(url)
    data_set=[]
    if not re.match(r"^https?",url)
    url="http://+url
    try:
    response =requests.get(url)
    exceptL

```

response=

sop=-999

filename=secure_filename(file.filename)

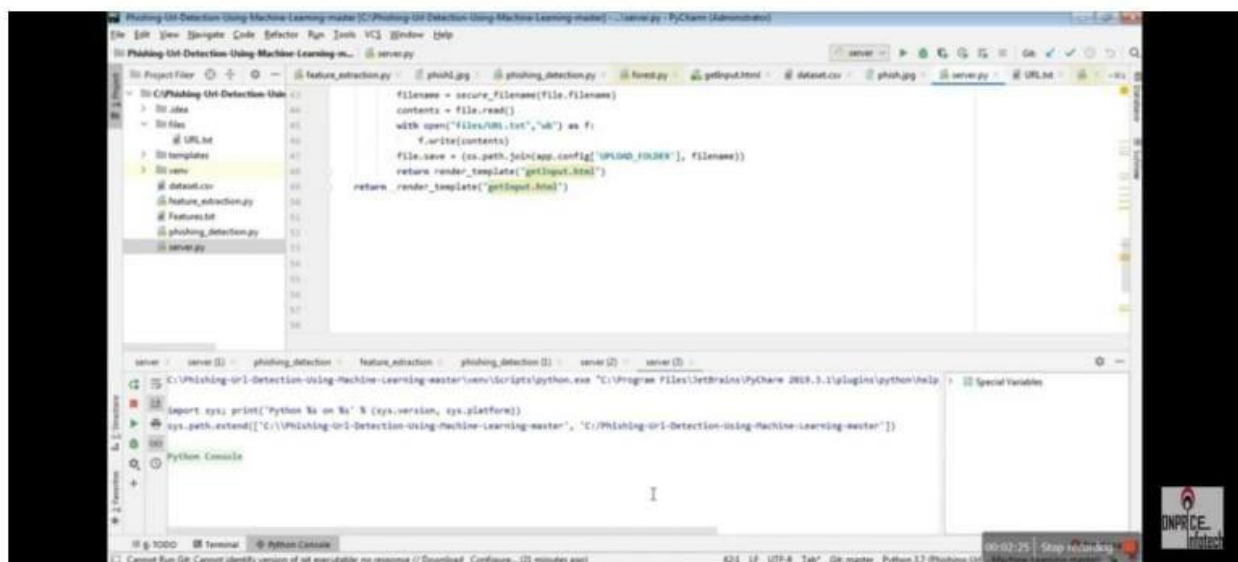
contents=file.read()

with open ("file/url.txt","ud")

return render_template("getInput.html")

return render_template("getInput.html")

sample output:



8. Testing:



Window 1



- 1) Web Templates
- 2) Site Cloner
- 3) Custom Import

99) Return to Webattack Menu

set:webattack>2

[~] Credential harvester will allow you to utilize the clone capabilities within SET
[~] to harvest credentials or parameters from a website as well as place them into a report
[~] This option is used for what IP the server will POST to.
[~] If you're using an external IP, use your external IP for this

set:webattack> IP address for the POST back in Harvester/Tabnabbing:192.168.0.192

[~] SET supports both HTTP and HTTPS

[~] Example: <http://www.thisisafakesite.com>

set:webattack> Enter the url to clone:facebook.com

[*] Cloning the website: <https://login.facebook.com/login.php>

[*] This could take a little bit...

The best way to use this attack is if username and password form fields are available. Regardless, this captures all POSTs on a website.

[*] The Social-Engineer Toolkit Credential Harvester Attack

[*] Credential Harvester is running on port 80

[*] Information will be displayed to you as it arrives below:

ATTENTION: A bind system call was requested on port: 80

The port has been changed. If connecting from outside GNURoot, use: 2080

```
File Actions Edit View Help


```

(preeti@kali):~$
$ cd Desktop
(preeti@kali):~/Desktop$
$ mkdir king-phisher
(preeti@kali):~/Desktop$
$ wget https://github.com/securestate/king-phisher/raw/master/tools/install.sh
--2022-04-08 20:50:17-- https://github.com/securestate/king-phisher/raw/master/tools/install.sh
Resolving github.com (github.com)... 13.224.219.28
Connecting to github.com (github.com)[13.224.219.28]:443... connected.
HTTP request sent, awaiting response... 301 Moved Permanently
Location: https://github.com/rsmusllp/king-phisher/raw/master/tools/install.sh [following]
--2022-04-08 20:50:20-- https://github.com/rsmusllp/king-phisher/raw/master/tools/install.sh
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.110.133, 185.199.109.133, 185.199.108.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)[185.199.110.133]:443... connected.
HTTP request sent, awaiting response... 302 found
Location: https://raw.githubusercontent.com/rsmusllp/king-phisher/master/tools/install.sh [following]
--2022-04-08 20:50:20-- https://raw.githubusercontent.com/rsmusllp/king-phisher/master/tools/install.sh
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.110.133, 185.199.109.133, 185.199.108.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)[185.199.110.133]:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 22801 (23K) [text/plain]
Saving to: 'install.sh'

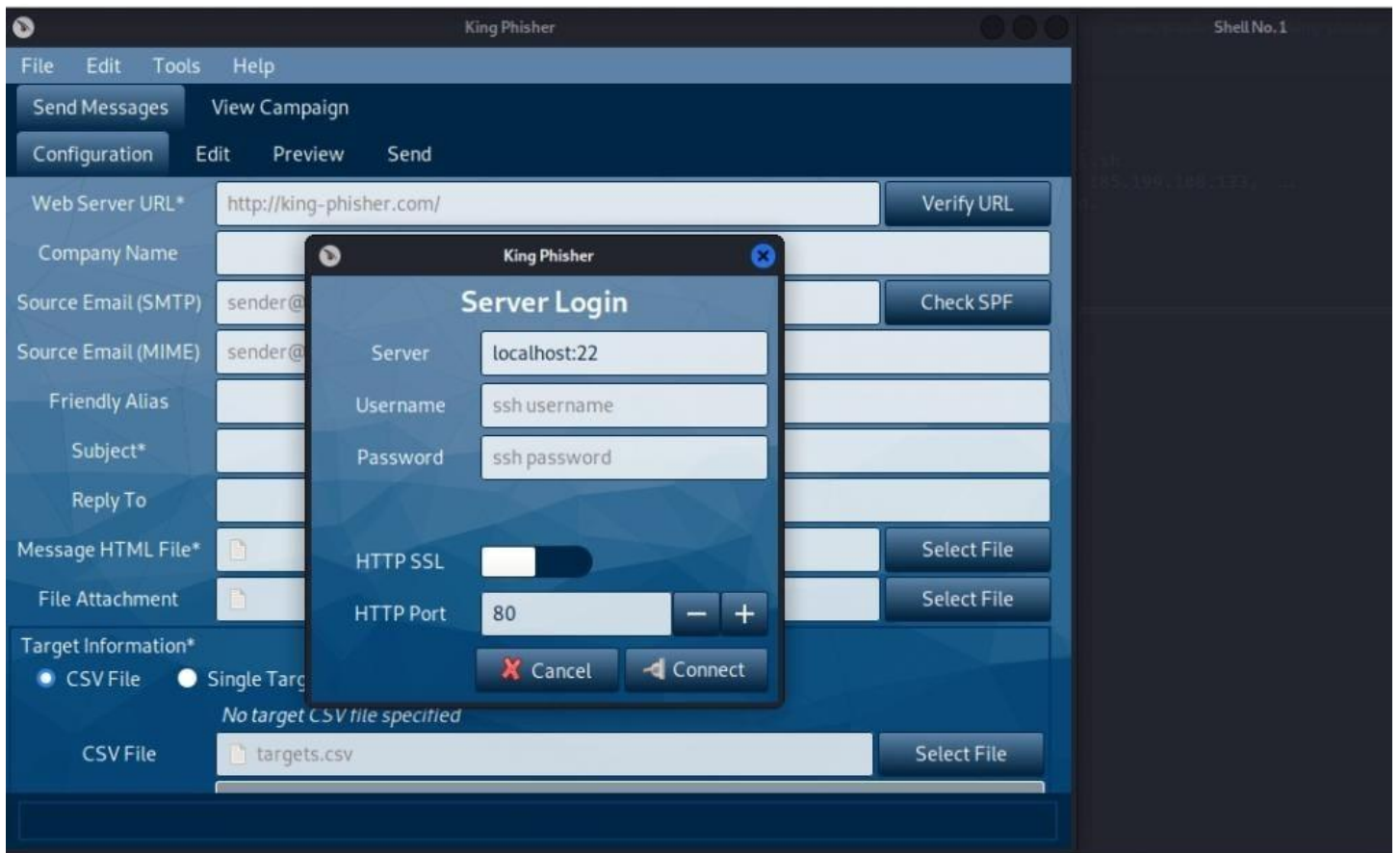
install.sh 100%[=====] 21.49K ---KB/s in 0s

2022-04-08 20:50:23 (64.3 MB/s) - 'install.sh' saved [22801/22801]

(preeti@kali):~/Desktop$
$ sudo bash ./install.sh
[sudo] password for preeti:
INFO: linux version detected as Kali
Install and use PostgreSQL? (Highly recommended and required for upgrading) [Y/n] █

```


```



9.RESULT:

Many users unwittingly click phishing domains every day and every hour. The attackers are targeting both the users and the companies. According to the 3rd Microsoft Computing Safer Index Report, released in February 2014, the annual worldwide impact of phishing could be very high as \$5 billion.



Connect with your friends
faster, wherever you are.
The Facebook application is available in
more than 2,500 phones.

Sign up
It's free (and all yours)

Sign up
Sign up

facebook

Dear Facebook user,

In an effort to make your online experience safer and more enjoyable, Facebook will be implementing a new login system that will affect all Facebook users. These changes will offer new features and increased account security.

Before you are able to use the new login system, you will be required to update your account.

Click [here](#) to update your account online now.

If you have any questions, reference our New User Guide.

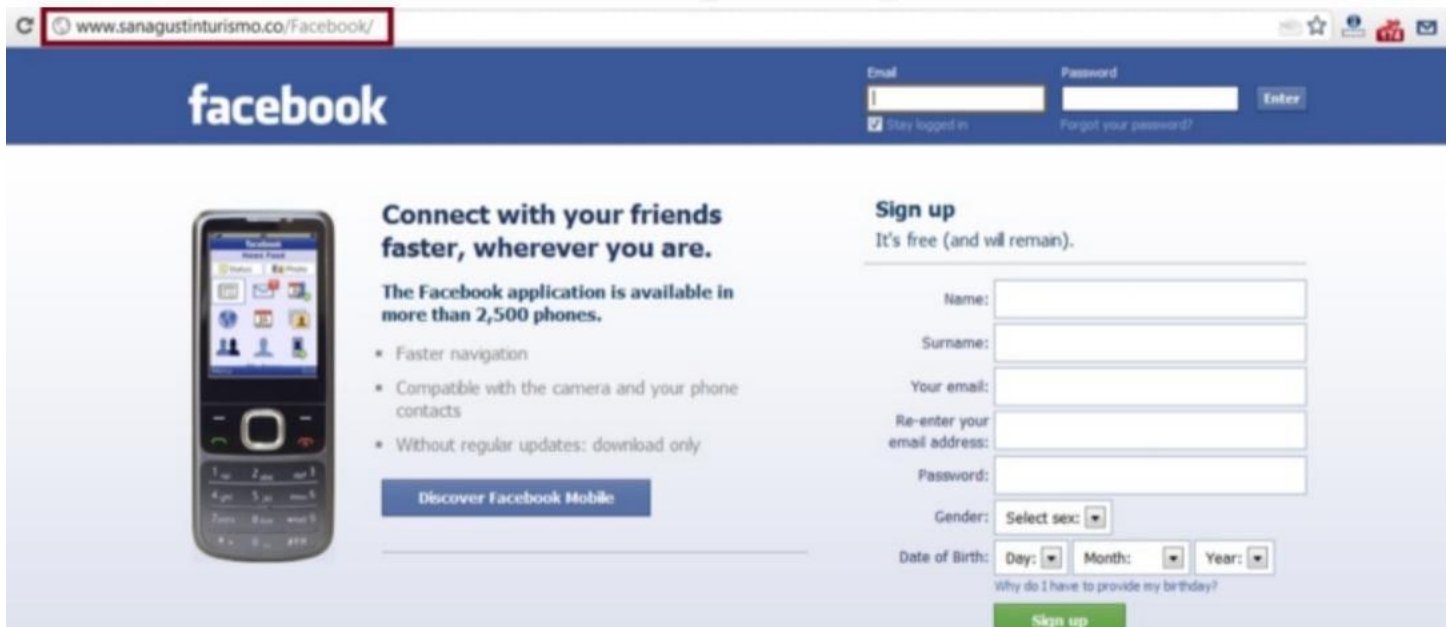
Thanks,
The Facebook Team

Update your
Facebook account

Update

attackers are targeting both the users
and the companies. According to the
3rd Microsoft Computing Safer Index

and the companies. According to the
3rd Microsoft Computing Safer Index



The screenshot shows the Facebook mobile app download page. At the top, the browser address bar displays 'www.sanagustinturismo.co/Facebook/'. The Facebook logo is on the left, and login fields for Email and Password are on the right. The main content area features a mobile phone displaying the Facebook app interface. To the right of the phone, the text reads: 'Connect with your friends faster, wherever you are. The Facebook application is available in more than 2,500 phones.' Below this, a list of features is provided: 'Faster navigation', 'Compatible with the camera and your phone contacts', and 'Without regular updates: download only'. A blue button labeled 'Discover Facebook Mobile' is positioned below the list. To the right of the phone, the 'Sign up' section is visible, stating 'It's free (and will remain).' and listing the required fields: Name, Surname, Your email, Re-enter your email address, Password, Gender (with a 'Select sex' dropdown), and Date of Birth (with Day, Month, and Year dropdowns). A green 'Sign up' button is at the bottom of the sign-up section.

www.sanagustinturismo.co/Facebook/

facebook

Email Password Enter

☒ Stay logged in [Forgot your password?](#)

Connect with your friends faster, wherever you are.

The Facebook application is available in more than 2,500 phones.

- Faster navigation
- Compatible with the camera and your phone contacts
- Without regular updates: download only

Discover Facebook Mobile

Sign up
It's free (and will remain).

Name:

Surname:

Your email:

Re-enter your email address:

Password:

Gender: Select sex:

Date of Birth: Day: Month: Year:

Why do I have to provide my birthday?

Sign up

The main reason is the lack of
awareness of users. But security
defenders must take precautions to
protect users from configuration errors

Open in app

Get started

Towards Data Science

Published in

Towards Data Science

Ebubekir Büber

Ebubekir Büber

Follow

Feb 8, 2018

.

11 min read

.

Listen

Save

Phishing URL Detection with ML

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 details.

Phishing is popular among attackers, since it is easier to trick someone into clicking a malicious link which seems legitimate than trying to break through a computer's defense systems. The malicious links within the body of the message are designed to make it appear that they go to the spoofed organization using that organization's logos and other legitimate contents.

In this article I explain: phishing domain (or Fraudulent Domain) characteristics, the features that distinguish them from legitimate domains, why it is important to detect these domains, and how they can be detected using machine learning and natural language processing techniques.

Many users unwittingly click phishing domains every day and every hour. The attackers are targeting both the users and the companies. According to the 3rd Microsoft Computing Safer Index Report, released in February 2014, the annual worldwide impact of phishing could be very high as \$5 billion.

What is the reason of this cost?

The main reason is the lack of awareness of users. But security defenders must take precautions to prevent users from confronting these harmful sites. Preventing these huge costs can start with making people conscious in addition to building strong security mechanisms which are able to detect and prevent phishing domains from reaching the user.

Characteristics of Phishing Domains

Lets check the URL structure for the clear understanding of how attackers think when they create a phishing domain.

Uniform Resource Locator (URL) is created to address web pages. The figure below shows relevant parts in the structure of a typical URL.

It begins with a protocol used to access the page. The fully qualified domain name identifies the server who hosts the web page. It consists of a registered domain name (second-level domain) and suffix which we refer to as top-level domain (TLD). The domain name portion is constrained since it has to be registered with a domain name Registrar. A Host name consists of a subdomain name and a domain name. An phisher has full control over the subdomain portions and can set any value to it. The URL may also have a path and file components which, too, can be changed by the phisher at will. The subdomain name and path are fully controllable by the phisher. We use the term FreeURL to refer to those parts of the URL in the rest of the article.

The attacker can register any domain name that has not been registered before. This part of URL can be set only once. The phisher can change FreeURL at any time to create a new URL. The reason security defenders struggle to detect

phishing domains is because of the unique part of the website domain (the FreeURL). When a domain detected as a fraudulent, it is easy to prevent this domain before an user access to it.

Some threat intelligence companies detect and publish fraudulent web pages or IPs as blacklists, thus preventing these harmful assets by others is getting easier. (cymon, firehol)

The attacker must intelligently choose the domain names because the aim should be convincing the users, and then setting the FreeURL to make detection difficult. Lets analyse an example given below.

Although the real domain name is active-userid.com, the attacker tried to make the domain look like paypal.com by adding FreeURL. When users see paypal.com at the beginning of the URL, they can trust the site and connect it, then can share their sensitive information to the this fraudulent site. This is a frequently used method by attackers.

Other methods that are often used by attackers are Cybersquatting and Typosquatting.

Cybersquatting (also known as domain squatting), is registering, trafficking in, or using a domain name with bad faith intent to profit from the goodwill of a trademark belonging to someone else. The cybersquatter may offer selling the domain to a person or company who owns a trademark contained within the name at an inflated price or may use it for fraudulent purposes such as phishing. For example, the name of your company is "abcompany" and you register as abcompany.com. Then phishers can register abcompany.net, abcompany.org, abcompany.biz and they can use it for fraudulent purpose.

Typosquatting, also called URL hijacking, is a form of cybersquatting which relies on mistakes such as typographical errors made by Internet users when inputting a website address into a web browser or based on typographical errors that are hard to notice while quick reading. URLs which are created with Typosquatting looks like a trusted domain. A user may accidentally enter an incorrect website address or click a link which looks like a trusted domain, and in this way, they may visit an alternative website owned by a phisher.

A famous example of Typosquatting is goggle.com, an extremely dangerous website. Another similar thing is youtube.com, which is similar to goggle.com except it targets Youtube users. Similarly, www.airfrance.com has been typosquatted as www.arifrance.com, diverting users to a website peddling discount travel. Some other examples; paywpal.com, microroft.com, applle.com, appie.com.

Features Used for Phishing Domain Detection

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.

URL-Based Features

Domain-Based Features

Page-Based Features

Content-Based Features

URL-Based Features

URL is the first thing to analyse a website to decide whether it is a phishing or not. As we mentioned before, URLs of phishing domains have some distinctive points. Features which are related to these points are obtained when the URL is processed. Some of URL-Based Features are given below.

Digit count in the URL

Total length of URL

Checking whether the URL is Typosquatted or not. (google.com → goggle.com)

Checking whether it includes a legitimate brand name or not (apple-icloud-login.com)

Number of subdomains in URL

Is Top Level Domain (TLD) one of the commonly used one?

Domain-Based Features

The purpose of Phishing Domain Detection is detecting phishing domain names. Therefore, passive queries related to the domain name, which we want to classify as phishing or not, provide useful information to us. Some useful Domain-Based Features are given below.

Its domain name or its IP address in blacklists of well-known reputation services?

How many days passed since the domain was registered?

Is the registrant name hidden?

Page-Based Features

Page-Based Features are using information about pages which are calculated reputation ranking services. Some of these features give information about how much reliable a web site is. Some of Page-Based Features are given below.

Global Pagerank

Country Pagerank

Position at the Alexa Top 1 Million Site

Some Page-Based Features give us information about user activity on target site. Some of these features are given below. Obtaining these types of features is not easy. There are some paid services for obtaining these types of features.

Estimated Number of Visits for the domain on a daily, weekly, or monthly basis

Average Pageviews per visit

Average Visit Duration

Web traffic share per country

Count of reference from Social Networks to the given domain

Category of the domain

Similar websites etc.

Content-Based Features

Obtaining these types of features requires active scan to target domain. Page contents are processed for us to detect whether target domain is used for phishing or not. Some processed information about pages are given below.

Page Titles

Meta Tags

Hidden Text

Text in the Body

Images etc.

By analysing these information, we can gather information such as;

Is it required to login to website

Website category

Information about audience profile etc.

All of features explained above are useful for phishing domain detection. In some cases, it may not be useful to use some of these, so there are some limitations for using these features. For example, it may not be logical to use some of the features such as Content-Based Features for the developing fast detection mechanism which is able to analyze the number of domains between 100.000 and 200.000. Another example would be, if we want to analyze new registered domains Page-Based Features is not very useful. Therefore, the features that will be used by the detection mechanism depends on the purpose of the detection mechanism. Which features to use in the detection mechanism should be selected carefully.

Detection Process

Detecting Phishing Domains is a classification problem, so it means we need labeled data which has samples as phish domains and legitimate domains in the training phase. The dataset which will be used in the training phase is a very important point to build successful detection mechanism. We have to use samples whose classes are precisely known. So it means, the samples which are labeled as phishing must be absolutely detected as phishing. Likewise the samples which are labeled as legitimate must be absolutely detected as legitimate. Otherwise, the system will not work correctly if we use samples that we are not sure about.

For this purpose, some public datasets are created for phishing. Some of the well-known one is PhishTank. These data sources are used commonly in academic studies.

Collecting legitimate domains is another problem. For this purpose, site reputation services are commonly used. These services analyse and rank available websites. This ranking may be global or may be country-based. Ranking mechanism depends on a wide variety of features. The websites which have high rank scores are legitimate sites which are used very frequently. One of the well-known reputation ranking service is Alexa. Researchers are using top lists of Alexa for legitimate sites.

When we have raw data for phishing and legitimate sites, the next step should be processing these data and extract meaningful information from it to detect fraudulent domains. The dataset to be used for machine learning must actually consist these features. So, we must process the raw data which is collected from Alexa, Phishtank or other data resources, and create a new dataset to train our system with machine learning algorithms. The feature values should be selected according to our needs and purposes and should be calculated for every one of them.

There so many machine learning algorithms and each algorithm has its own working mechanism. In this article, we have explained Decision Tree Algorithm, because I think, this algorithm is a simple and powerful one.

Initially, as we mentioned above, phishing domain is one of the classification problem. So, this means we need labeled instances to build detection mechanism. In this problem we have two classes: (1) phishing and (2) legitimate.

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.

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.

Now, the most important question about Decision Trees is not answered yet. The question is that which feature will be located as the root? and which ones must come after the root? Choosing features intelligently effects efficiency and success rate of algorithms directly.

So, how does decision tree algorithm select features?

Decision Tree uses a information gain measure which indicates how well a given feature separates the training examples according to their target classification. The name of the method is Information Gain. The mathematical equation of information gain method is given below.

High Gain score means that the feature has a high distinguishing ability. Because of this, the feature which has maximum gain score is selected as the root. Entropy is a statistical measure from information theory that characterizes (im-)purity of an arbitrary collection S of examples. The mathematical equation of Entropy is given below.

Original Entropy is a constant value, Relative Entropy is changeable. Low Relative Entropy Score means high purity, likewise high Relative Entropy Score means low purity. As we move down the tree, we want to increase the purity, because high purity on the leaf implies high success rate.

In the training phase, dataset is divided into two parts by comparing the feature values. In our example we have 14 samples. “+” sign representing phishing class, and “-” sign representing legitimate class. We divided these samples into two parts according to the length feature. Seven of them settle right, the other seven of them settle left. As shown in the figure below, right part of tree has high purity, so it means low Entropy Score (E), likewise left part of tree has low purity and high Entropy Score (E). All calculations were done according to the equations given above. Information Gain Score about the length feature is 0,151.

The Decision Tree Algorithm calculates this information for every feature and selects features with maximum Gain scores. To growth the tree, leaves are changed as a node which represents a feature. As the tree grows downwards, all leaves will have high purity. When the tree is big enough, the training process is completed.

The Tree created by selecting the most distinguishing features represents model structure for our detection mechanism. Creating mechanism which has high success rate depends on training dataset. For the generalization of system success, the training set must be consisted of a wide variety of samples taken from a wide variety of data sources. Otherwise, our system may working with high success rate on our dataset, but it can not work successfully on real world data. BNN

GitHub link:

<https://github.com/IBM-EPBL/IBM-Project-40742-1660633765>

Demo link:

<https://drive.google.com/file/d/1k-DrGIUcFn3Mjg14t2Ijz5tfMS6p-RDM/view?usp=drivesdk>