CSM601 Mini project 2B
Detection system for Phishing website

By

TE_B_49 Rohan Shinde

TE_B_55 Pranjal Sonawane

TE_B_58 Pranay Tate

TE_B_60 Gaurav Thakur

under the guidance of

Ms. Snehal Shinde

Department of Computer Engineering

Pillai HOC College of Engineering and Technology,Rasayani

University of Mumbai, Mumbai

A.Y 2022-2023

Content

- Introduction
- Objectives
- Literature review
- Limitations of existing system
- Implemented System
- Results
- Conclusion
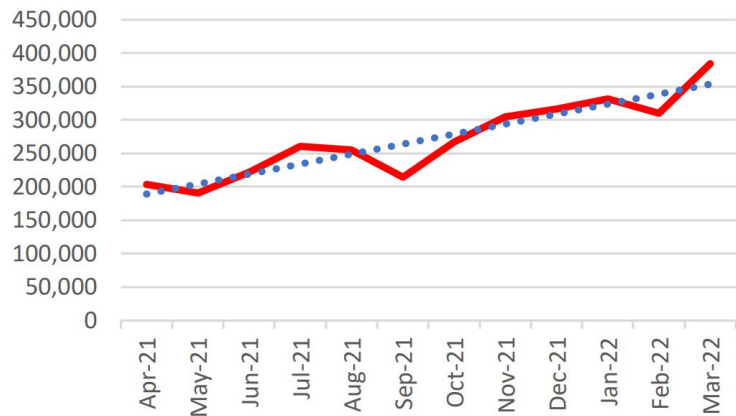- Reference

Introduction

What is phishing ?

When it comes to cyberattacks, phishing remains as one of the ever-popular techniques use for malicious intents. Phishing is a type of attack which combines technical tactic and social engineering in an attempt to obtain personal or otherwise sensitive information such as login credentials, credit card details, or company secrets from the victim.
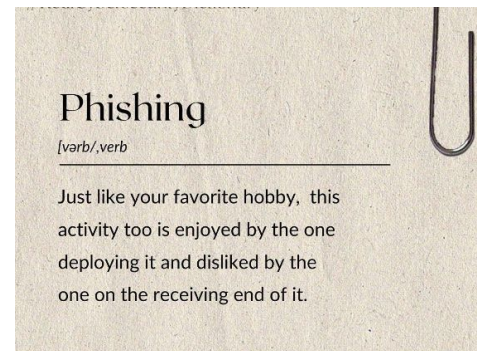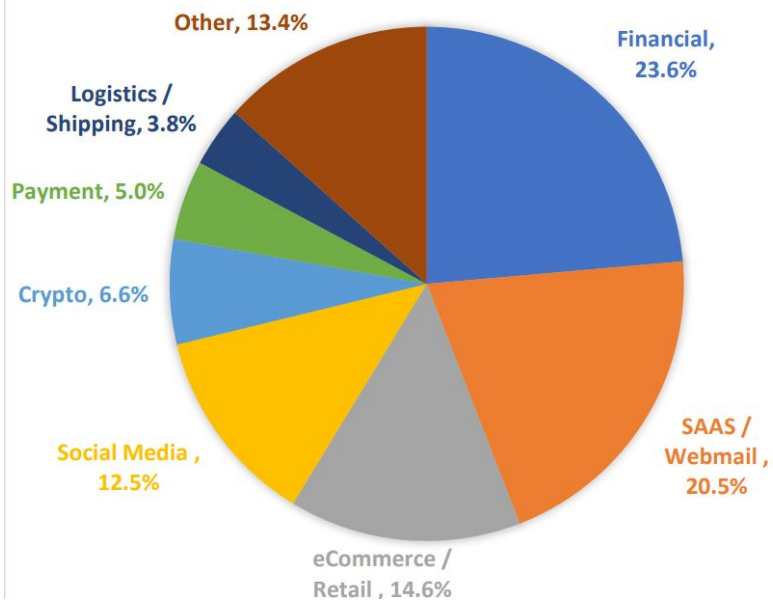
Anti-Phishing Working Group (APWG) Report

- 1,025,968 + phishing attacks.
- The financial sector was the most frequently victimized by phishing in Q1, with 23.6% of all attacks. Attacks against SaaS and webmail providers continued to be numerous.
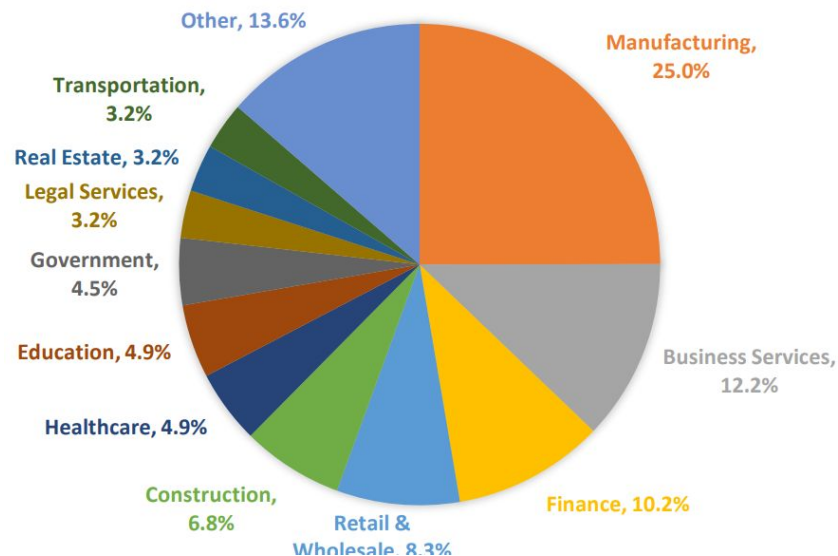- Phishing against cryptocurrency targets inched up to 6.6% of attacks

Phishing Attacks, 2Q2021 - 1Q2022

MOST-TARGETED INDUSTRIES, 1Q2022

- Financial, 23.6%
- SAAS / Webmail, 20.5%
- eCommerce / Retail, 14.6%
- Social Media, 12.5%
- Crypto, 6.6%
- Payment, 5.0%
- Logistics / Shipping, 3.8%
- Other, 13.4%

RANSOMWARE VICTIM INDUSTRIES, 1Q2022

- Manufacturing, 25.0%
- Business Services, 12.2%
- Finance, 10.2%
- Retail & Wholesale, 8.3%
- Construction, 6.8%
- Healthcare, 4.9%
- Education, 4.9%
- Government, 4.5%
- Legal Services, 3.2%
- Real Estate, 3.2%
- Transportation, 3.2%
- Other, 13.6%

Objective

Our project deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Various ML algorithms are used to detect phishing websites. Aim of the project is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm

Literature Survey

Name of paper	Author	Technique Used	Limitation
Phishing Detection: Analysis of Visual Similarity Based Approaches	Ankit Kumar Jain and B. B. Gupta	Visual Appearance	If the phishing website is partially copied (less than 50%) from the legitimate website, then none of the visual similarity based approach can detect it.
Utilisation of website logo for phishing detection	Wei KingTiong	website logo	slow down the feature extraction steps and not reliable
Anti-phishing based on automated individual white-list	Weili Han , Yueran Le	Maintain a white-list of user's all familiar Login User Interfaces (LUIs) of web sites	Not reliable for daily use

Implemented System

- Detect and block phishing website manually
- Block phishing email by various spam filter and software
- Installing anti-phishing , antivirus software in system
- Checking of SSL certificate
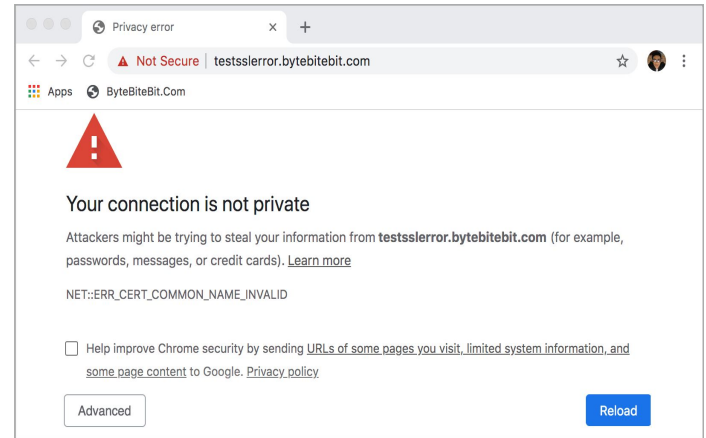
Proposed System

- Collect dataset containing phishing
- Extract the required features
- Exploratory analysis
- Feature engineering
- Model Training , Testing
- Evaluation
- Deployment

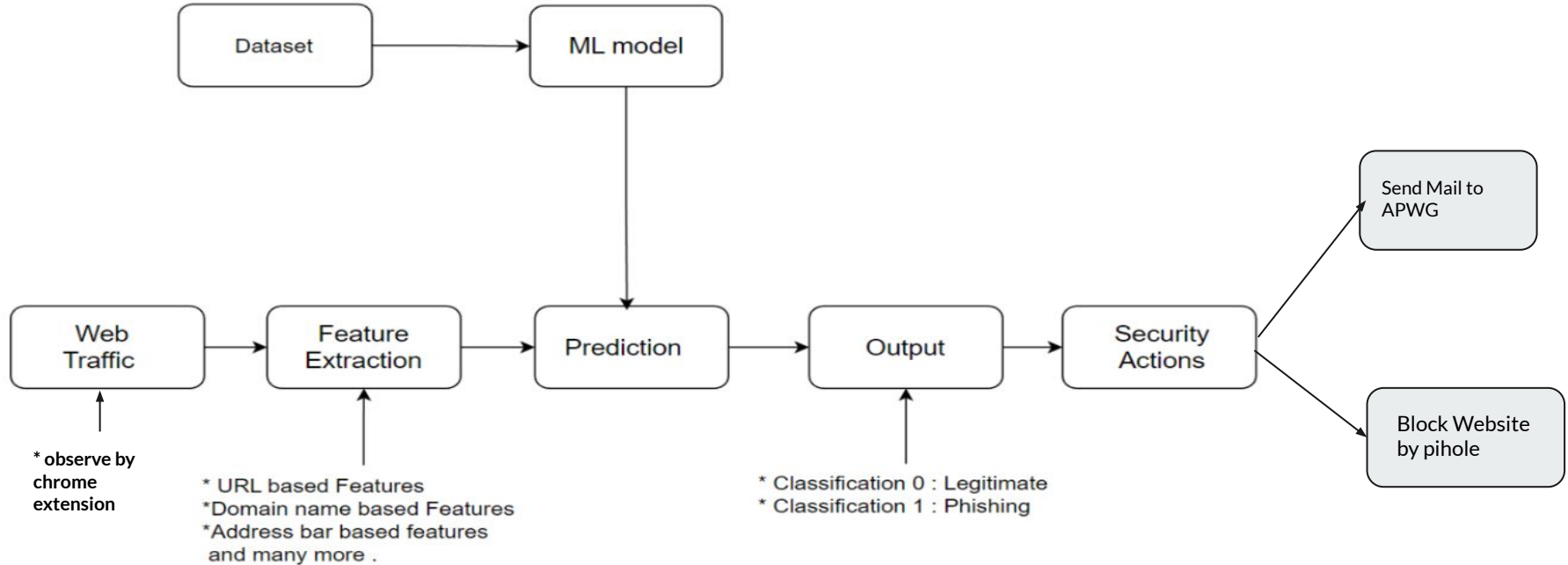# FEATURE EXTRACTION

Classification output: 0 = legitimate, 1 = phishing

1. having_IP_Address
2. URL_Length
3. Shortening_Service
4. having_At_Symbol
5. double_slash_redirecting
6. Prefix_Suffix
7. having_Sub_Domain
8. SSL_State

9. Domain_registeration_length
10. Favicon
11. Open_ports
12. HTTPS_token_in_URL
13. Request_URL
14. URL_of_Anchor
15. Links_in_tags
16. Server_Form_Handler

17. Submitting_to_email
18. Abnormal_URL
19. Site_Redirect
20. on_mouseover_changes
21. RightClick_Disabled
22. popUpWindow
23. Iframe_redirection
24. age_of_domain

25. DNS_Record
26. web_traffic_rank
27. Page_Rank
28. Google_Index
29. Links_pointing_to_page
30. Statistical_report - top phishing domains

Architecture

Dataset → ML model

Web Traffic → Feature Extraction → Prediction → Output → Security Actions → Send Mail to APWG

Security Actions → Block Website by pihole

* observe by chrome extension

* URL based Features
*Domain name based Features
*Address bar based features
 and many more .

* Classification 0 : Legitimate
* Classification 1 : Phishing

Problem we saw

- Most of the phishing attack came from the targeted ads .
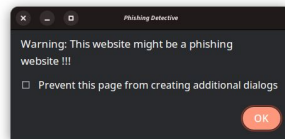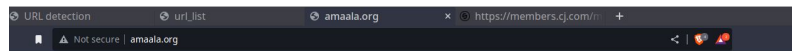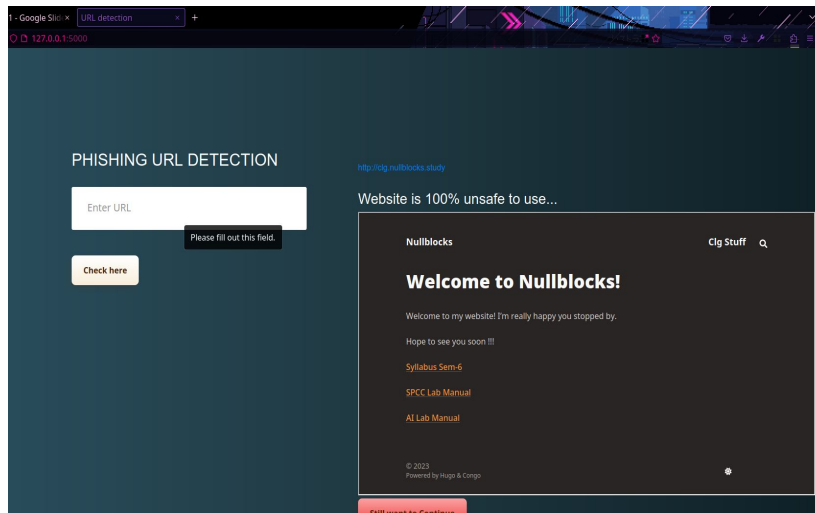- Ads are everywhere .

Solution

Pi-Hole

- Network-wide protection
- Block in-app advertisements
- Improve network performance
- Monitor statistics
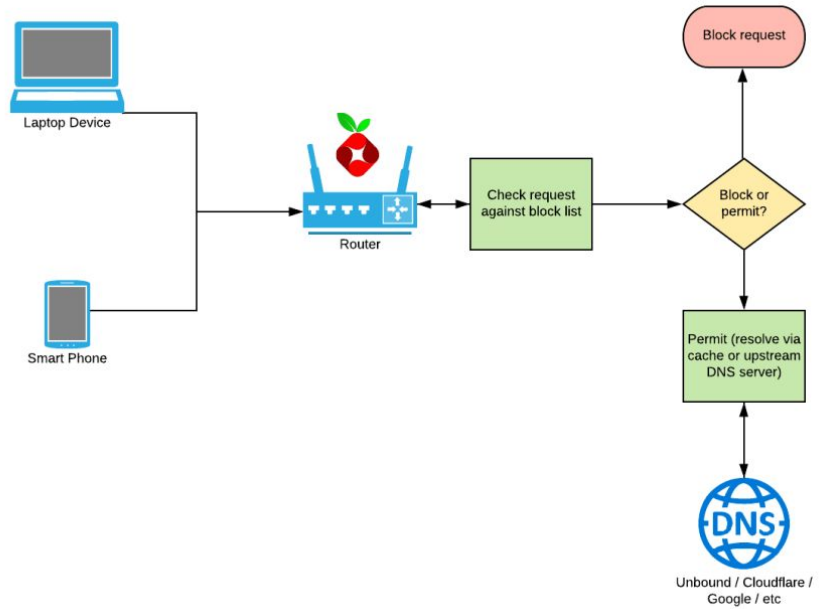
Results

Conclusion

In conclusion, our project aimed to protect users from phishing websites by identifying and alerting them to potential threats, while the use of pihole significantly reduced the risk of users being exposed to malicious ads that could lead to phishing attacks.

Reference

- https://www.sciencedirect.com/science/article/pii/S0167404815001145
- https://dl.acm.org/doi/abs/10.1145/1456424.1456434
- https://www.hindawi.com/journals/scn/2017/5421046/
- https://www.researchgate.net/publication/340695779_Detecting_Phishing_Website_Using_Machine_Learning
- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0258361
- https://www.irjet.net/archives/V5/i3/IRJET-V5I3580.pdf

Thank You