Detection of Phishing Websites Hosted on Free Web Hosting Domains Using Machine Learning

Under the supervision of
Dr. Meenakshi Tripathi

Rohit Lohar - 2019UCP1373
Sneh Aashish Gupta - 2019UCP1900
Divya Rathore - 2019UCP1370

MNIT, JAIPUR

May 12, 2023

Table Of Contents

Introduction

- Phishing is a type of cyber attack that involves using fraudulent emails, text messages, or other forms of electronic communication to trick individuals into providing sensitive information, such as passwords, credit card numbers, or other personal data.
- some of the facts about phishing attacks in 2023 :
  - One of the main reasons for data breaches is phishing. The average cost of a data breach rose from $4.24m in 2021 to $4.35m in 2022, according to IBM's 2022 Cost of Data Breach Report.
  - According to Verizon's analysis from 2022, phishing was implicated in 36% of all data breaches. By 2022, one ransomware or phishing assault was predicted to happen every 11 seconds
  - The top four most commonly spoofed tech companies are Amazon (13%), Google (8%), Facebook (9%), and Apple (2%) in that order.

Motivation

- Phishing attacks based on FHDs are increasing day by day.
- A large-scale study of these attacks shows that phishing websites hosted on FHDs stay online at least 1.5 times longer than normal phishing URLs.
- On average, they have 1.7 times less coverage from anti-phishing blocklists than regular phishing attacks.
- Their coverage time is also 3.8 times slower, and anti-phishing tools only catch half as many of them.
- The hosting site took an average of 12.2 hours to get rid of only 23.6% of FHD URLs a week after they first appeared
- Past studies have used machine learning to detect phishing problems.

Objectives

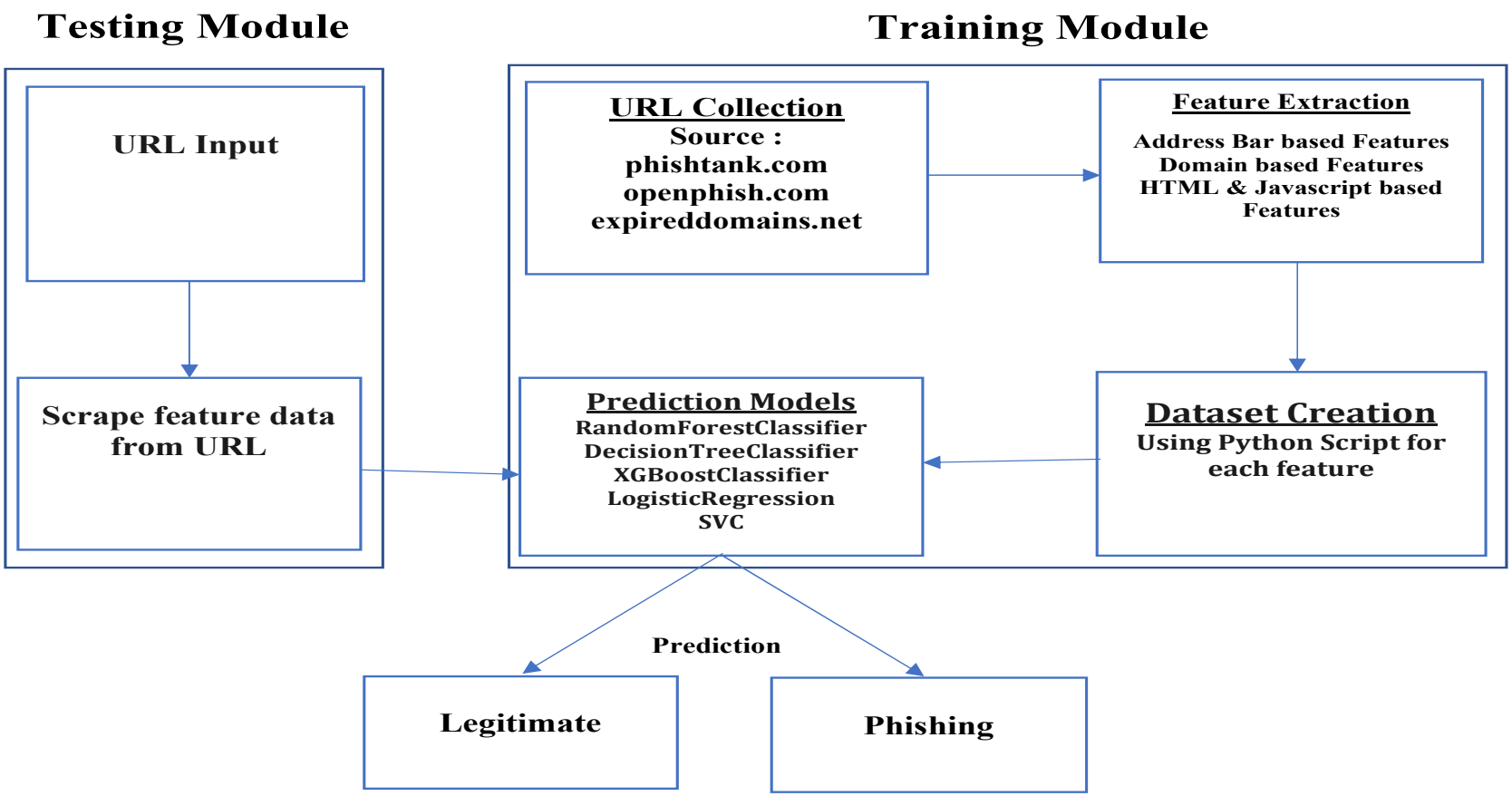
Detection of phishing websites hosted on free web hosting domains (FHDs) using machine learning techniques

- to identify phishing websites hosted on free web hosting domains.
- to come up with trustworthy features and creation of the up-to-date dataset.
- to develop machine learning models to precisely detect phishing websites on FHD.

Related Works

- Shirazi et al.[6] used 6,000 legitimate websites and 6,000 phishing websites to make their Fresh-Phish dataset.Using Gaussian kernel SVM, they got a 90% success rate for the fresh phishing test data.
- Panigua et al.[5] used LightGBM model to train their self made 60k dataset using 30 features to get 94.7% accuracy.
- Roy et al.[4] found and categorized a new group of phishing sites that take advantage of free hosting services.They collected 2764 FHD URLs and their dataset with 7 features.Using their model's Random Forest Classifier, they were able to achieve an accuracy of 97.30% in their classifications.

Proposed Methodology



- URL Collection
- Features Extraction
- Dataset Creation

URL Collection

- A total of 50,917 phishing websites from phishtank.com and 499 phishing websites from openphish.com are collected and supposed to be phishing.
- We collected a total of 8,19,599 websites from expireddomains.net and supposed to be legitimate.

Features Extraction

- 32 features
- The collected features are grouped into the following groups:
  - Address Bar based Features
  - Domain based Features
  - HTML and Javascript based Features

Table: Features Extracted

HTTPS Scheme	DNS Record	Favicon	Non standard port
'@' Symbol	Age of Domain	Copyright Symbol	Sub pages count
URL Length	Domain End	Copyright year	Indexed pages count
URL Depth	Iframe	Domain name around copyright	MLT tags ratio
Redirection	Mouse Over	Link Ratio	Visibility mode
HTTPS present in Domain	Right Click	SSL name and domain name	Anchor URL
Prefix Suffix by '-'	Web Forwarding	Server Form Handler	Request URL
Sub Domain Count	Google Index	Form mail	Empty URL

Dataset Creation

- Since our study deals with Free Hosted Domains based URL,so we filtered this from our URL collections by string matching FHD providers name with URL's domain name.
- The free domain providers which we used in this study is given below in the Table.

Table: Free Hosted Domain Providers

weebly	duckdns	000webhost	blogspot	wix	googlesites
github	firebase	squareup	zohoforms	wordpress	googleforms
sharepoint	yolasite	myftp	godaddysites	mailchimp	atwebpages
glitch	webnode	herokuapp	website	netlify	hpage
infinityfree	byethost	hyperphp	awardspace	freehostia	freehosting
freewebhostingarea	hostpapa	ultahost	porkbun	bluehost	googiehost
x10hosting	freehosting	freehostia	siteground	dreamhost	hostgator
wordpress	domainracer	freehostingnoads	freehostingnoads	freewebhostingarea	namecheap
inmotionhosting	a2hosting	interserver			

Cont...

- For just confirmations, we used the Virus Total tool to check these websites. Virus Total is a web-based service that analyzes suspicious files and URLs.
- Therefore, after applying filters, we were successful in obtaining the following number of URLs:
  - Legitimate FHD URLs: 2499
  - Phishing FHD URLs: 2423
- We were also interested in exploring the possibility of training and testing our features on non-FHD URLs.
  - Legitimate Non-FHD URLs: 2499
  - Phishing Non-FHD URLs : 2511

Performance Evaluation

- we carried out two separate sets of experiments.On each dataset, we trained and tested six distinct machine learning models.
- We examined the performance metrics of the machine learning model that provided the best accuracy for each dataset.
- These experiments are as follows:
  - Training and testing machine learning models on non-FHD dataset
  - Training and testing machine learning models on FHD dataset
- We splitted our dataset into 80:20 ratio for train and test respectively.

Training and testing machine learning models on non-FHD dataset

- We trained our dataset on 6 ML models and got the results as folows :

Table: Accuracy on Non-FHD Dataset

Model Name	Train Accuracy	Test Accuracy
Random Forest Classifier	0.999	0.977
XG Boost Classifier	0.989	0.975
Multilayer Perceptrons(MLP)	0.986	0.97
Decision Tree Classifier	0.976	0.97
Logistic Regression	0.97	0.968
Support Vector Machine	0.951	0.941

- Performance Metrics of non-FHD dataset on Random Forest Classifier

Table: Performance Metrics on non-FHD dataset

Model Name	Accuracy	Precision	Recall	F1-Score
Random Forest Classifier	0.977	0.996	0.958	0.977

Comparison of the proposed methodology with existing methodology Panigua et al.[5]

Parameters / Methodology	Model name	No. of features	size of dataset	Accuracy	Precision	Recall	F1-score
PANIAGUA et al.2017 [7]	LightGBM Classifier	**38**	**60000**	0.947	0.953	0.939	0.946
Proposed Method	Random Forest Classifier	32	5010	**0.977**	**0.996**	**0.958**	**0.977**

Figure: Comparison of Existing Methodology with Proposed Methodology

Training and testing machine learning models on FHD dataset

- We trained our dataset on 6 ML models and got the results as follows:

Table: Accuracy on FHD Dataset

Model Name	Train Accuracy	Test Accuracy
Multilayer Perceptrons(MLP)	0.972	0.965
Random Forest Classifier	0.982	0.955
XG Boost Classifier	0.965	0.947
Support Vector Classifier	0.932	0.937
Logistic Regression	0.919	0.93
Decision Tree Classifier	0.982	0.927

- Performance Metrics of FHD dataset on Multilayer Perceptrons(MLP)

Table: Performance Metrics on FHD dataset

Model Name	Accuracy	Precision	Recall	F1-Score
Multilayer Perceptrons(MLP)	0.965	0.950	0.981	0.966

Comparison of the proposed methodology with existing methodology Roy et al.[4]

Parameters / Methodology	Model name	No.ofFHD providers	No. of features	size of dataset	Accuracy	Precision	Recall	F1-score
Roy et al.2022[1]	Random Forest Classifier	24	7	2764	**0.973**	**0.975**	0.975	0.975
Proposed Method	Multilayer Percep-tron	**50**	**32**	**4922**	0.965	0.950	**0.981**	**0.996**

Figure: Comparison of Existing Methodology with Proposed Methodology

Conclusions

- We developed an up-to-date dataset of 4922 FHD URLs that were just recently collected.
- Using a multilayer perceptron neural network with 32 features, we got the best accuracy of 96.5% , covering the maximum characters of phishing websites.
- We also put our features to the test on a set of self-made non-FHD URL datasets and got a 97.7 % success rate.

Future works

- we can extract and utilize the visual characteristics of phishing and legitimate websites and use convolutional neural networks to detect phishing websites using visual features.
- We can also come up with new adaptive features that can detect evolving phishing attacks.

References

1 S. K. Deval, M. Tripathi, B. Bezawada, and I. Ray, "x-phish: Days of future past" ‡: Adaptive privacy preserving phishing detection," in 2021 IEEE Conference on Communications and Network Security (CNS). IEEE, 2021, pp. 227–235.

2 M. KAYTAN and D. HANBAY, "Effective classificationof phishing web pages based on new rules by using extreme learning machines," Computer Science, vol. 2, no. 1, pp. 15–36, 2017.

3 R. M. Mohammad, F. Thabtah, and L. McCluskey, "Phishing websites features," School of Computing and Engineering, University of Huddersfield, 2015.

4 S. S. Roy, U. Karanjit, and S. Nilizadeh, "A large-scale analysis of phishing websites hosted on free web hosting domains," arXiv preprint arXiv:2212.02563, 2022.

5 M. S anchez-Paniagua, E. F. Fern andez, E. Alegre, W. Al-Nabki, and V. Gonzalez-Castro, "Phishing url detection: A real-case scenario through login urls," IEEE Access, vol. 10, pp. 42 949–42 960, 2022.

References

6 H. Shirazi, K. Haefner, and I. Ray, "Fresh-phish: A framework for auto-detection of phishing websites," in 2017 IEEE international conference on information reuse and integration (IRI). IEEE, 2017, pp. 137–143.

7 "Virustotal," https://www.virustotal.com/gui/home/url, accessed: 2023-04-26.

8 "Top most intriguing recent phishing attacks statistics," https://www.getastra.com/blog/security-audit/phishing-attack accessed: 2023-04-24.

9 "Insufficient phishing websites data," https://zvelo.com/single-use-phishing-urls-need-zero-second-detection/, accessed: 2023-04-26.

10 "Url collections from phishtank," http://data.phishtank.com/data/onlinevalid.csv , accessed: 2023-04-05.

References

11 "Url collections from openphish," https://openphish.com/feed.txt , accessed: 2023-04-05.

12 "Url collections from expireddomains," https://www.expireddomains.net/alexa-top-websites/, accessed: 2023-04-05.

13 "Source of roc curve," https://en.wikipedia.org/wiki/Receiveroperatingcharacteristic, accessed: 20230426.

14 "Source of random forest tutorials," https://towardsdatascience.com/random-forests-algorithm-explained-with-a-real-life-example-and-some-python-code-affbfa5a942c, accessed:2023-04-26.

15 "Source of xgboost tutorials," https://towardsdatascience.com/a-beginners-guide-to-xgboost-87f5d4c30ed7, accessed: 2023-04-26.

References

16 "Ibm's 2022 cost of data breach statistics," https://www.egress.com/blog/phishing/phishing-statistics- round-up, accessed: 2023-04-24.

17 "Verizon 2022 data report," https://www.verizon.com/business/en-gb/resources/reports/dbir/, accessed: 2023-04-24.

18 "Unifying the global response to cybercrime report 2022," https://www.comparitech.com/blog/vpn-privacy/phishing-statistics-facts/, accessed: 2023-04-24.

19 "Phishing report by comparitech 2022," https://www.comparitech.com/blog/vpn-privacy/phishing-statistics-facts/, accessed: 2023-04-24.

20 "Decision tree theory concept," https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm, accessed: 2023-04-26.

References

21 "Logistic regression theory concept," https://www.javatpoint.com/logistic-regression-in-machine-learning, accessed: 2023-04-26.

22 "Support vector machine theory concept," https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm, accessed: 2023-04-26.

23 "Multilayer perceptron theory concept," https://medium.com/codex/introduction-to-how-an-multilayer-perceptron-works-but-without-complicated-math-a423979897ac, accessed: 2023-04-26.