



Feature Selection for Machine Learning-based Phishing Websites Detection

Smriti Dangwal & Arghir-Nicolae Moldovan

School of Computing, National College of Ireland, Mayor Street, IFSC, Dublin 1, Ireland smriti.dangwal@student.ncirl.ie; arghir.moldovan@ncirl.ie

International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA 2021) Virtual Conference

> Updated for NCI Research Day 25th June 2021

Outline

Introduction

- Motivation and Goals
- Methodology
- Results
- Conclusions
- **Q&A**



Introduction (1)

Phishing

- A type of social engineering attack
- Commonly used to deceive users to reveal sensitive information such as login credentials or credit card details
- Also used to deploy malicious software like ransomware
- May start with an e-mail or text message
- May ask the user to visit a URL

Some phishing statistics

C-MRIC

National

College

Ireland

- Over 2.11 million phishing websites detected by Google in 2020 (AtlasVPN, 2020)
- Increasing number of unique phishing websites and email subjects detected in 2020 (Anti Phishing Working Group, 2000)
- 220% increase in phishing attacks in 2020, with many using certificates with "covid" or "corona" in name (F5 Labs, 2020)





Source: https://apwg.org/trendsreports/



Introduction (2)

Phishing websites

C-MRIC

National

College of

Ireland

- Often target well-known brands
- □ Are increasingly realistic
- Need for both increased user training/awareness and automatic phishing detection solutions



Sign In	
What is your	e-mail address?
My e-mai	il address is:
Do you have	an Amazon.com password?
O No, I a	am a new customer.

OpenPhish

4

Global Phishing Activity

The Global Phishing Activity provides real-time insight into live phishing pages that were observed by OpenPhish. The data on this page is updated every five minutes with information from the past 24 hours period.

List of identified brands (updated monthly)



Top 10 Targeted E	Brands	Top 10 Sector	rs	Top 10 ASNs				
Office365	8.6%	Financial	31.8%	AS46606 Unified Layer	9.5%			
Facebook, Inc.	8.6%	Online/Cloud Service	18.9%	AS15169 Google LLC	8.6%			
Amazon.com Inc.	7.3%	Social Networking	13.1%	AS204915 Hostinger I	5.4%			
Outlook	6.4%	e-Commerce	9.5%	AS13335 Cloudflare, I	5.1%			
Tencent	4.2%	Email Provider	8.1%	AS8100 QuadraNet E	5.1%			
Credit Agricole S.A.	3.9%	Telecommunications	6.2%	AS27647 Weebly, Inc.	4.3%			
Chase Personal Banki	3.8%	Payment Service	4.9%	AS22612 Namecheap,	4.0%			
La Banque postale	3.3%	Logistics & Couriers	3.2%	AS16509 Amazon.co	3.1%			



Introduction (3)

Blacklist approach

- Relies on the phishing websites to be listed/known
- Well known blacklists include Google Safe Browsing, OpenPhish, PhishTank
- May involve real users that can report phishing websites, verify them

Phis	ihTanl	C [®] Out of the Net, into	the Tank.	username <mark>Register</mark> <u>For</u>	got Password	Sig
Home Ad	ld A Phish Ver	ify A Phish Phish Search	1 Stats FAQ Develop	ers Mailing Lists	My Account	
Stats					Monthly Stats Archive:	
Online, va 10,15	lid phishes 56	Total Submissions 6,964,005	Total Votes 21,630,691	Daily Phisl	nes Submitted	
Phishes Ve	erified as Valid	Suspected	Phishes Submitted		-4	
Total:	2,894,583	Total:	6,963,939	1500		
Online:	10,179	Online:	15,063			-
Offline:	2,884,404	Offline:	6,948,876	1000 -		
Most Ac	tive Users (c	out of 187,273 total))	500 -		
1 cle	anmx	3,080,604 ph	ishes			
2 <u>Phi</u>	shReporter	1,216,400 ph	ishes	y 25	y 29	n 16
2 ant	inhishing	105 503 phie	hes	May	AMA A A A A A A A A A A A A A A A A A A	



URL:		
	I'm not a robot	reCAPTCHA Privacy - Terms
Additional details about the phishing		//
(Optional)	Submit Report	Google



Motivation and Goals

Machine learning-based phishing website detection

- Aims to build models that detect phishing websites based on other characteristics (e.g., URL, content)
- Complement blacklist approach (e.g., if the website is not blacklisted)

Prior research

- Some prior research works that applied machine learning used a high number of features
- May not be feasible to extract some features for real-time detection
- U While some works compared ML algorithms on multiple datasets, they did not combine the datasets

Research Goals

- Perform feature selection for building robust machine learning-based phishing website detection models
- Identify common features between different website phishing datasets
- □ Investigate the usefulness of Variance Inflation Factor (VIF) as a feature selection method



Methodology (1)

Systematic approach

Follows the KDD methodology for knowledge discovery and data mining

Data selection

- Two datasets with 30 and 48 features
- DS1-30 contains both

C-MRIC

National

College of

Ireland

- Internal features (i.e., derived from webpage URL and HTML/JavaScript source code)
- External features (i.e., obtained from querying third party services such as DNS, search engine, WHOIS records, etc.)
- DS2-48 only contains internal features



Source: Costagliola et al. (2009)

Dataset Code	Feature Category Feature Examples						
	URL Based	having_IP_Address, URL_Length,					
		HTTPS_token, etc.					
	Abnormal	Request_URL, URL_of_Anchor,					
	Based	Links_in_tags, etc.					
031-30	HTML/JS	Redirect, on_mouseover, RightClick,					
	Based	Based popUpWidnow, etc.					
	Domain Based	DNSRecord, web_traffic, Page_Rank,					
	Domain Baseu	Google_Index, etc.					
	URL Based	NumDots, UrlLength, AtSymbol, etc.					
	Abnormal	AbnormalExtFormActionR,					
DS2-48	ADHOTHAI	ExtMetaScriptLinkRT, etc.					
	HTML/JS	RightClickDisabled, ExtFavicon,					
	Based	PopUpWindow, etc.					



Methodology (2)

Data Preparation and Transformation

- Datasets were clean
- DS2-48 had one attribute with all values 0
- 18 common features were identified
- DS2-18 data was transformed to match the binary {-1, 1} or categorical {-1, 0, 1} format used by DS1-18

Dataset Code	Number Instances	Phishing Class	Legitimate Class	# Categorical Features	# Numeric Features
DS1-30	11055	11 20/	EE 70/	30	0
DS1-18	11055	44.5%	55.7%	18	0
DS2-48	10000	F.00/	F 00/	29	19
DS2-18	10000	50%	50%	11	7
DS12-18	21055	170/	E 20/	18	0
DS12-13	21022	4/%	55%	13	0

DS1-18, DS12-18	DS2-18	DS12-13
having_IP_Address	IpAddress	
having_Sub_Domain	SubdomainLevel *	\checkmark
Links_pointing_to_page	PctExtHyperlinks *	\checkmark
Submitting_to_email	SubmitInfoToEmail	\checkmark
double_slash_redirecting	DoubleSlashInPath	\checkmark
URL_Length	UrlLength *	\checkmark
Favicon	ExtFavicon	\checkmark
Prefix_Suffix	NumDashInHostname *	\checkmark
SFH	AbnormalFormAction	\checkmark
Iframe	IframeOrFrame	\checkmark
having_At_Symbol	AtSymbol	
SSLfinal_State	NoHttps	\checkmark
on_mouseover	FakeLinkInStatusBar	
URL_of_Anchor	PctNullSelfRedirectHyperlinks *	\checkmark
popUpWidnow	PopUpWindow	
Request_URL	PctExtResourceUrls *	\checkmark
RightClick	RightClickDisabled	
Links_in_tags	ExtMetaScriptLinkRT *	\checkmark

Note: * indicates numeric features, \checkmark indicates selected features.



8)

Methodology (3)

Feature Selection

- p-value analysis was used to test the significance of independent features
- Spearman rank-order correlation was used to test for collinearity between pairs of features
- Variance inflation factor (VIF) was used to identify multicollinearity (i.e., collinearity between three or more features even if no pair of variables has a particularly high correlation)

$$\mathrm{VIF}_i = rac{1}{1-R_i^2}$$

- \square R_i^2 is the coefficient of determination from a multiple regression model that predicts the *i*-th feature based on all other features
- Feature selection is performed by removing all features with a VIF score of 5 and above which indicate critical multicollinearity issues (Hair et al., 2019)
- 13 features were selected for DS12-13

C-MRIC

National

College of

Ireland

having_IP_Address	1.00	0.15	0.25	0.44	0.32	0.03	0.26	0.07	-0.00	0.29	-0.00	0.31	0.08	0.17	0.07	0.21	-0.10	-0.29	0.06
URL_Length	0.15	1.00	0.07	0.06	0.35	0.12	0.22	0.02	0.10	0.20	0.00	0.47	0.06	0.07	0.06	0.10	-0.09	-0.11	-0.08
having_At_Symbol	0.25	0.07	1.00	0.15	0.21	0.02	0.17	0.22	0.01	0.18	-0.06	0.21	0.29	0.32	0.21	0.35	0.06	-0.07	0.02
louble_slash_redirecting	0.44	0.06	0.15	1.00	0.16	0.03	0.10	0.03	-0.03	0.14	-0.09	0.17	0.04	0.14	0.04	0.13	-0.07	-0.18	-0.04
Prefix_Suffix	0.32	0.35	0.21	0.16	1.00	0.24	0.51	0.02	-0.01	0.58	0.02	0.58	0.03	0.18	0.10	0.23	-0.24	-0.14	0.12
having_Sub_Domain	0.03	0.12	0.02	0.03	0.24	1.00	0.31	-0.02	0.02	0.27	0.09	0.23	0.02	0.04	0.07	0.05	-0.06	-0.02	0.17
SSLfinal_State	0.26	0.22	0.17	0.10	0.51	0.31	1.00	0.00	0.11	0.57	0.10	0.46	0.04	0.13	0.05	0.14	-0.16	-0.11	0.45
Favicon	0.07	0.02	0.22	0.03	0.02	-0.02	0.00	1.00	0.13	0.01	-0.21	0.02	0.38	0.50	0.26	0.66	0.29	-0.15	0.03
Request_URL	-0.00	0.10	0.01	-0.03	-0.01	0.02	0.11	0.13	1.00	0.03	-0.28	0.04	0.02	-0.00	-0.04	-0.01	0.14	-0.22	0.11
URL_of_Anchor	0.29	0.20	0.18	0.14	0.58	0.27	0.57	0.01	0.03	1.00	0.13	0.46	0.01	0.16	0.11	0.19	-0.17	0.01	0.39
Links_in_tags	-0.00	0.00	-0.06	-0.09	0.02	0.09	0.10	-0.21	-0.28	0.13	1.00	-0.01	-0.05	-0.06	0.00	-0.09	-0.09	0.26	0.20
SFH	0.31	0.47	0.21	0.17	0.58	0.23	0.46	0.02	0.04	0.46	-0.01	1.00	0.07	0.18	0.07	0.23	-0.21	-0.19	0.01
Submitting_to_email	0.08	0.06	0.29	0.04	0.03	0.02	0.04	0.38	0.02	0.01	-0.05	0.07	1.00	0.41	0.27	0.48	0.24	-0.09	-0.15
on_mouseover	0.17	0.07	0.32	0.14	0.18	0.04	0.13	0.50	-0.00	0.16	-0.06	0.18	0.41	1.00	0.41	0.73	0.25	-0.08	0.01
RightClick	0.07	0.06	0.21	0.04	0.10	0.07	0.05	0.26	-0.04	0.11	0.00	0.07	0.27	0.41	1.00	0.37	0.31	-0.09	0.03
popUpWidnow	0.21	0.10	0.35	0.13	0.23	0.05	0.14	0.66	-0.01	0.19	-0.09	0.23	0.48	0.73	0.37	1.00	0.22	-0.15	-0.02
Iframe	-0.10	-0.09	0.06	-0.07	-0.24	-0.06	-0.16	0.29	0.14	-0.17	-0.09	-0.21	0.24	0.25	0.31	0.22	1.00	-0.11	-0.11
Links_pointing_to_page	-0.29	-0.11	-0.07	-0.18	-0.14	-0.02	-0.11	-0.15	-0.22	0.01	0.26	-0.19	-0.09	-0.08	-0.09	-0.15	-0.11	1.00	0.12
Result	0.06	-0.08	0.02	-0.04	0.12	0.17	0.45	0.03	0.11	0.39	0.20	0.01	-0.15	0.01	0.03	-0.02	-0.11	0.12	1.00
	ŝ	÷	0	ŋ	.×	. <u>c</u>	e	Ę		r	s	т	lie	Ŀ	×	≥	ē	Ð	ŧ
	dres	engi	/mp	ectin	Suff	oma	Stat	Ivico	Ъ [–]	ncha	taç	SF	ema	eove	tolic	oup	fram	pag	Sesu
	PA		, S	edire	efix	ŏ	na	Ц	lest	₹_A	s_in		ф	snoi	Righ	Wd	<u> </u>	р	ш.
		UR	9_A	ц Г	Pre	Sub	SLfi		Sequ	۲_ ۲	Link		ting.	u u		Dqo		Iting	
	wing		avin	slas		ing	S		ш.	IJ			bmit	ō		đ		poir	
	ha		Ĺ	ble		hav							Su						
				nop														Lir	



-1.00

double

-0.75

-0.50

-0.25

0.00

0.25

0.50

0.75

1.00

Methodology (4)

🗆 Data Mining

Two ML algorithms were selected for building binomial classification models

Random Forest (RF)

- □ Was shown to outperform a variety of other ML algorithms in many previous studies
- Tends to perform well using default settings
- Gradient Boosting Machine (GBM)
 - □ Was often not included in comparison by previous studies
 - Requires more hyperparameter tuning
- □ The DRF and GBM algorithm implementations from the H2O v3 open-source framework were used
- Models were built using different sets of hyperparameter values to identify optimal values

Evaluation

- Data was split into training and test set with 80:20 ratio
- Performance metrics computed: accuracy, precision, recall/sensitivity, specificity, AUC

Methodology Workflow

Implementation

Best model built with 13 features was integrated into a Python application that takes a URLs as input, extracts features from the live website, and predicts if it is legitimate or phishing





S. Dangwal & A.-N. Moldovan – "Feature Selection for Machine Learning-based Phishing Websites Detection", CyberSA 2021, Virtual Conference, June 2021

Results (1)

Model Performance

- All models achieved over 92% accuracy
- DRF and GBM models perform very close
 - DRF models have slightly higher accuracy than GBM models for five datasets
 - □ GBM model has higher accuracy for DS2-48
- Good baseline performance
 - DS1-30: DRF accuracy of 0.974
 - DS2-48: DRF accuracy of 0.985
- 18 common features
 - \square DS1-18: DRF accuracy of $\textbf{0.952} \rightarrow \textbf{0.022}$ decrease from baseline
 - □ DS2-18: DRF accuracy of **0.937** \rightarrow **0.048** decrease from baseline
 - Higher drop in performance for DS2-18 can be explained by the data transformation of DS2-18 features to match the DS1-18 format
 - DS12-18: DRF accuracy of 0.937, AUC of 0.985
- 13 optimal features

C-MRIC

National

College of

Ireland

DS12-13: DRF accuracy of 0.937, AUC of 0.979





Results (2)

Comparison with Previous Research

- Second experiment compared the performance of DRF and GBM algorithms with best results achieved by previous research papers on the DS1-30 and DS2-48 datasets
- Used the same validation techniques and data split ratio as reported by the authors of those papers
- □ DS1-30: DRF model achieved **lower** accuracy than previous works
- DS1-48: DRF model achieved **higher** accuracy than previous works

Reference	Dataset	Data Split	ML Algorithm	Acc.
Subasi et al. [10]		10f-CV	RF	0.974
Our results	DS1-30	10f-CV	DRF	0.973
Rahman et al. [9]		65:35%	ERT	0.970
Our results		65:35%	DRF	0.967
Rahman et al. [9]	DS2-48	65:35%	ERT	0.980
Our results		65:35%	DRF	0.981
Chiew et al. [17]		70:30%	RF	0.962
Our results		70:30%	DRF	0.984



Results (3)

College of

Ireland

Model Building Time

DRF building time on DS12-13 is 40% lower than the time on DS1-30 and 56% lower than on DS2-48

Feature Extraction Time

- A smaller DS3 dataset of live websites was used
- **~6.5** times lower extraction time for legitimate e-mails and **~3.5** times lower for phishing e-mails



Conclusions

- Performed feature selection to build a more robust machine learning model for phishing website detection
- Two datasets with 30 and 48 features were selected and analysed to identify 18 matching features
- Feature selection using variance inflation factor was conducted to identify 13 optimal features
- □ RF performs very well for phishing detection
 - □ 13 features model achieved 0.937 accuracy and 0.979 AUC
 - □ Results confirm prior research findings, but more hyperparameter tuning may be required for GBM
- □ Future work will focus on comparing VIF with other feature selection methods



Thank you for your attention!



