National College of Ireland

C-MRiC CENTRE FOR MULTIDISCIPLINARY RESEARCH, INNOVATION AND COLLABORATION ®

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

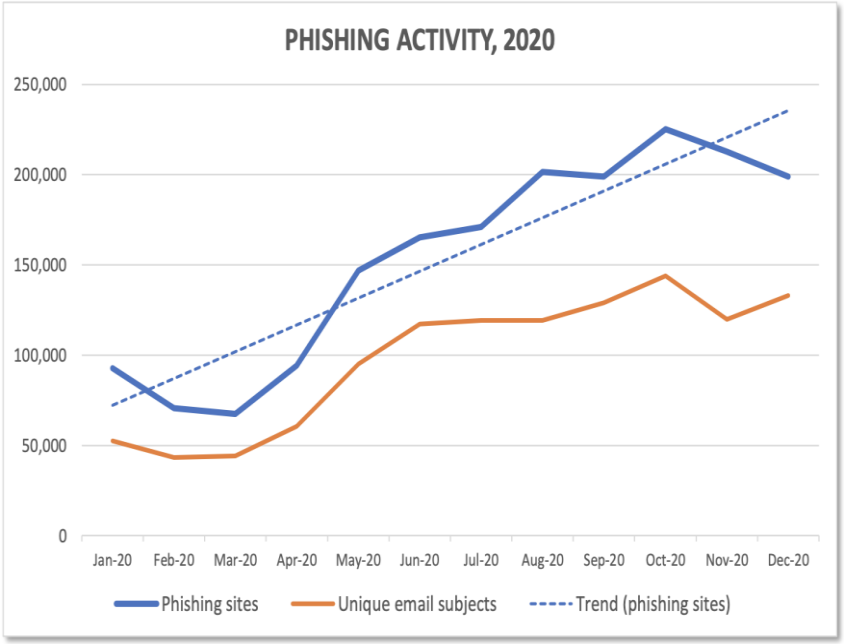
Outline

- 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
  - 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)

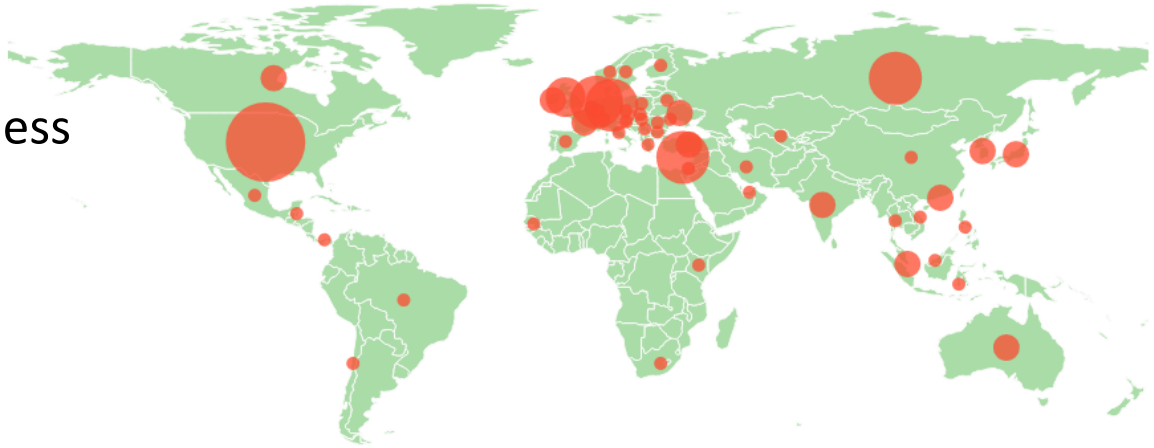
- Phishing websites
  - Often target well-known brands
  - Are increasingly realistic
  - Need for both increased user training/awareness and automatic phishing detection solutions

OpenPhish / Phishing Feeds / IP Reputation Feed / Global Phishing Activity

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 Brands

Office365	8.6%
Facebook, Inc.	8.6%
Amazon.com Inc.	7.3%
Outlook	6.4%
Tencent	4.2%
Credit Agricole S.A.	3.9%
Chase Personal Banki...	3.8%
La Banque postale	3.3%

Top 10 Sectors

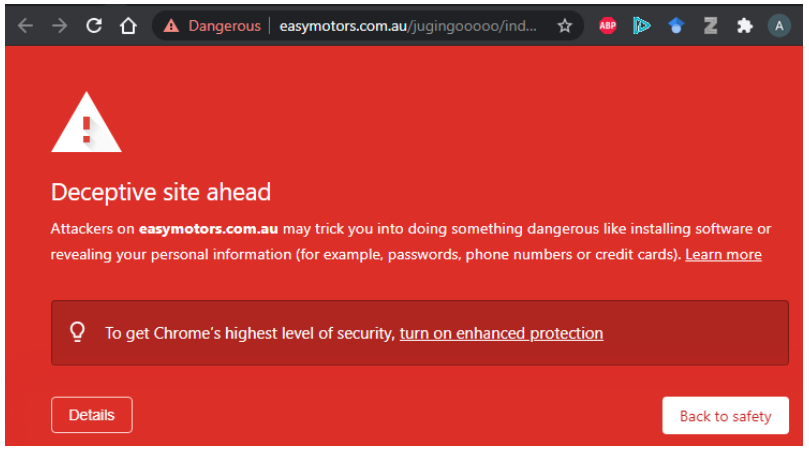
Financial	31.8%
Online/Cloud Service	18.9%
Social Networking	13.1%
e-Commerce	9.5%
Email Provider	8.1%
Telecommunications	6.2%
Payment Service	4.9%
Logistics & Couriers	3.2%

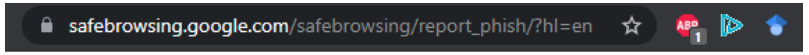
Top 10 ASNs

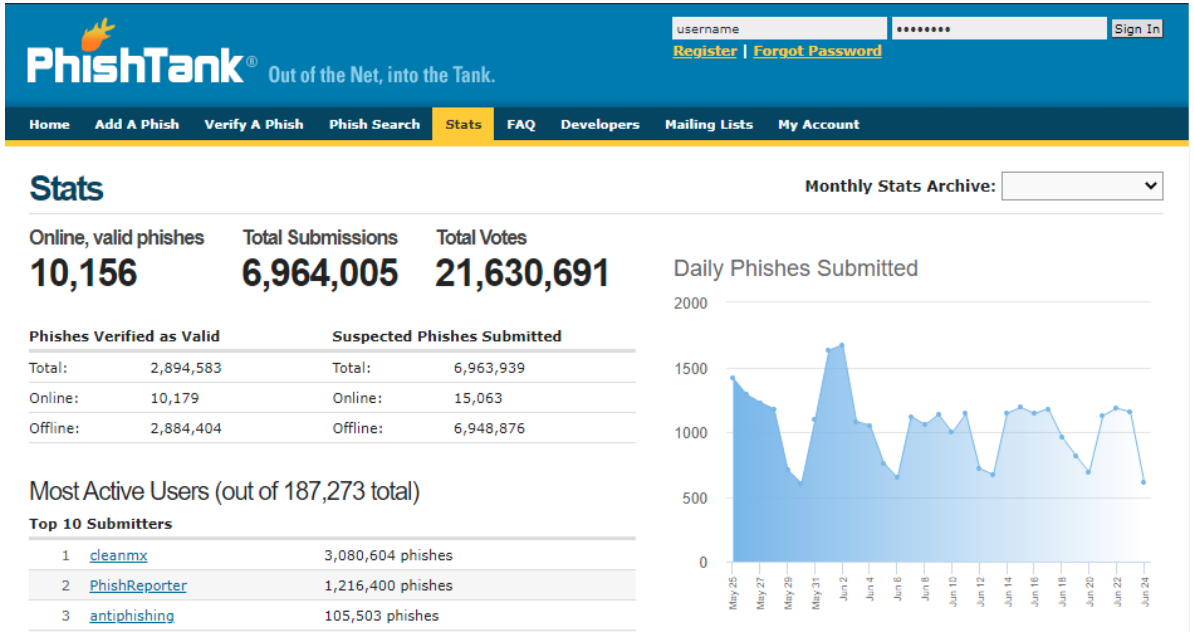
AS46606 Unified Layer	9.5%
AS15169 Google LLC	8.6%
AS204915 Hostinger I...	5.4%
AS13335 Cloudflare, I...	5.1%
AS8100 QuadraNet E...	5.1%
AS27647 Weebly, Inc.	4.3%
AS22612 Namecheap,...	4.0%
AS16509 Amazon.co...	3.1%

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







Motivation and Goals

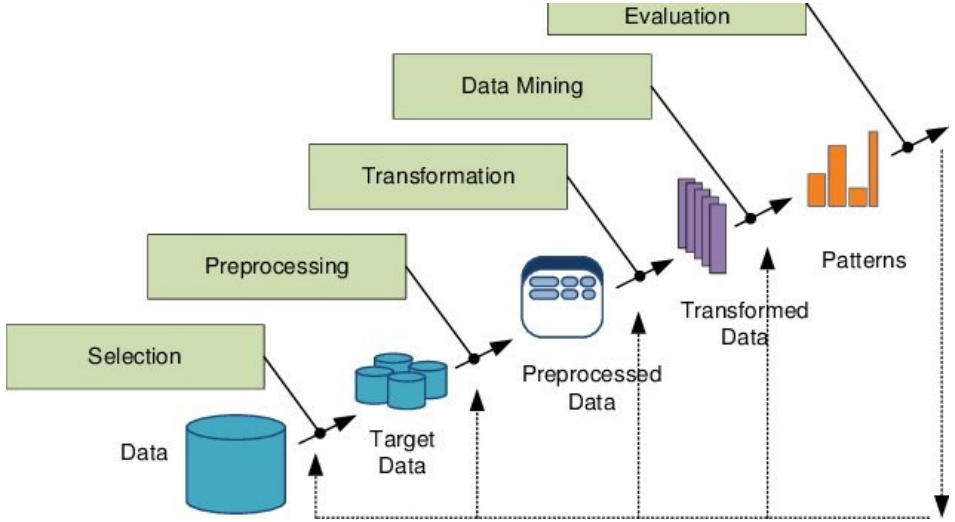
- Machine learning-based phishing website detection
  - Aims to build models that detect phishing websites based on other characteristics (e.g., URL, content)
  - Complement blacklist approach (e.g., if the website is not blacklisted)

- 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
  - 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



Source: Costagliola et al. (2009)

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

Dataset Code	Feature Category	Feature Examples
DS1-30	URL Based	having_IP_Address, URL_Length, HTTPS_token, etc.
	Abnormal Based	Request_URL, URL_of_Anchor, Links_in_tags, etc.
	HTML/JS Based	Redirect, on_mouseover, RightClick, popUpWidnow, etc.
	Domain Based	DNSRecord, web_traffic, Page_Rank, Google_Index, etc.
DS2-48	URL Based	NumDots, UrlLength, AtSymbol, etc.
	Abnormal	AbnormalExtFormActionR, ExtMetaScriptLinkRT, etc.
	HTML/JS Based	RightClickDisabled, ExtFavicon, 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	44.3%	55.7%	30	0
DS1-18				18	0
DS2-48	10000	50%	50%	29	19
DS2-18				11	7
DS12-18	21055	47%	53%	18	0
DS12-13				13	0

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

Note: * indicates numeric features, ✓ indicates selected features.

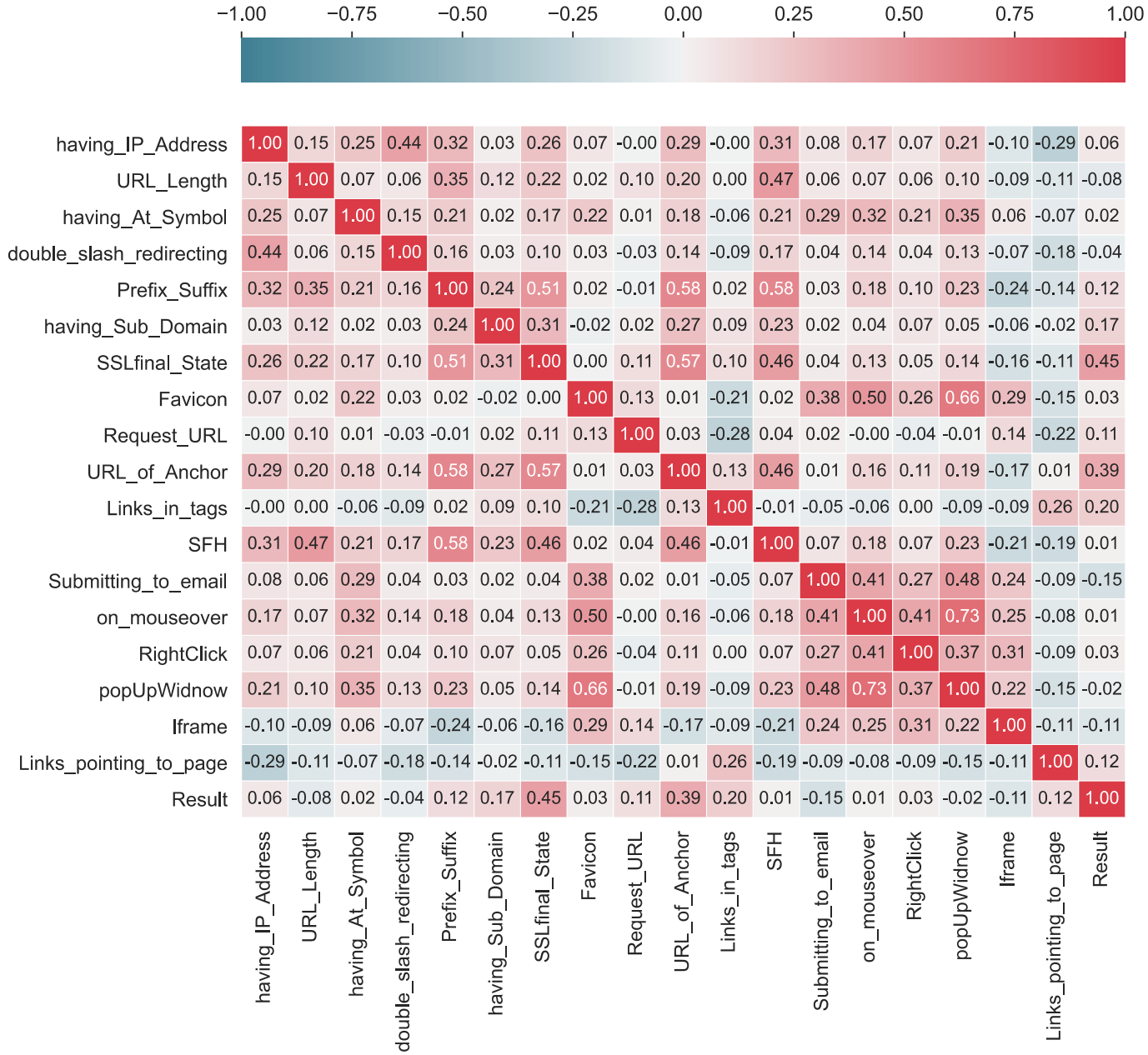
Methodology (3)

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)

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

- 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



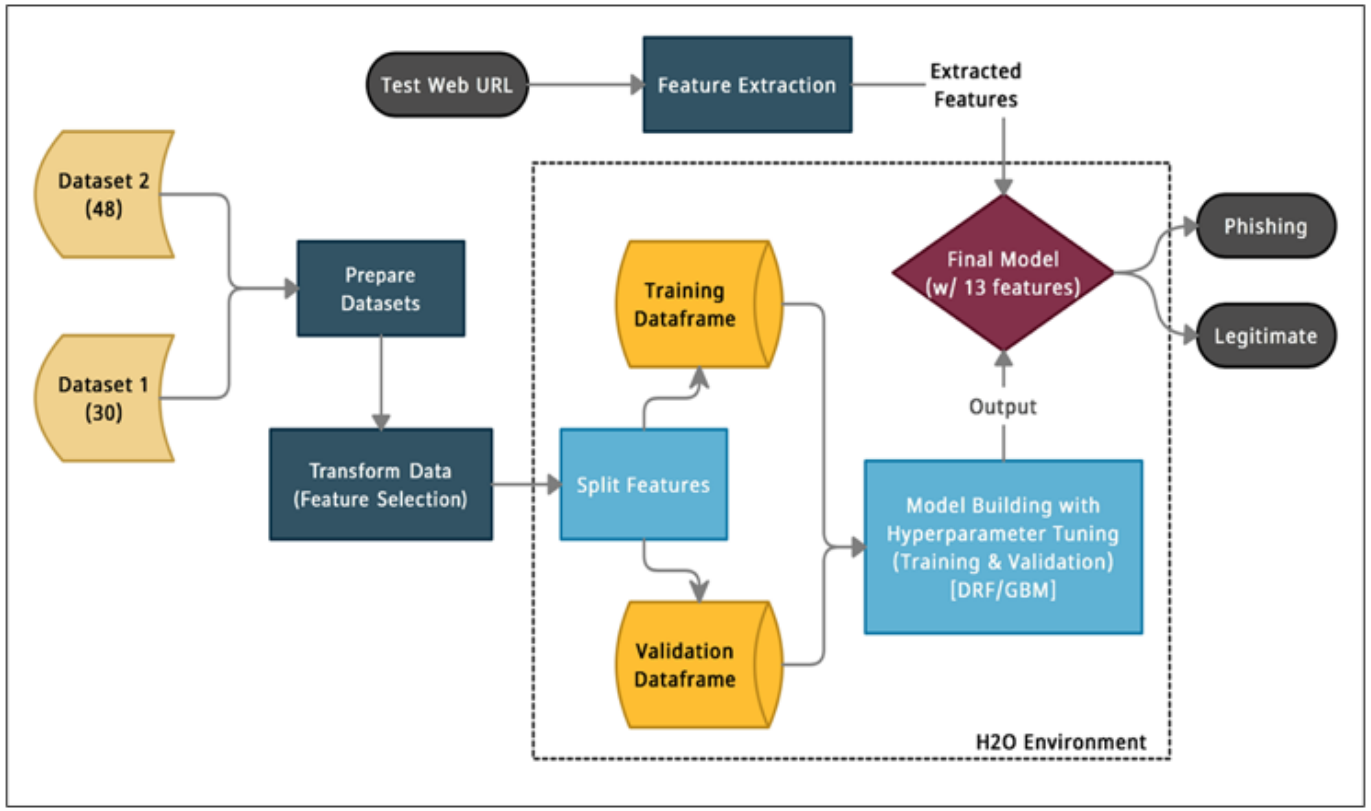
Methodology (4)

- 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

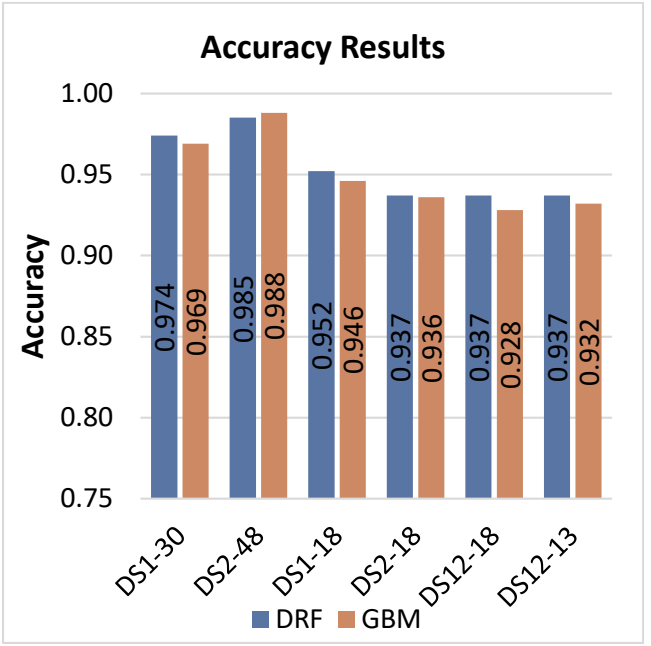
Methodology Workflow

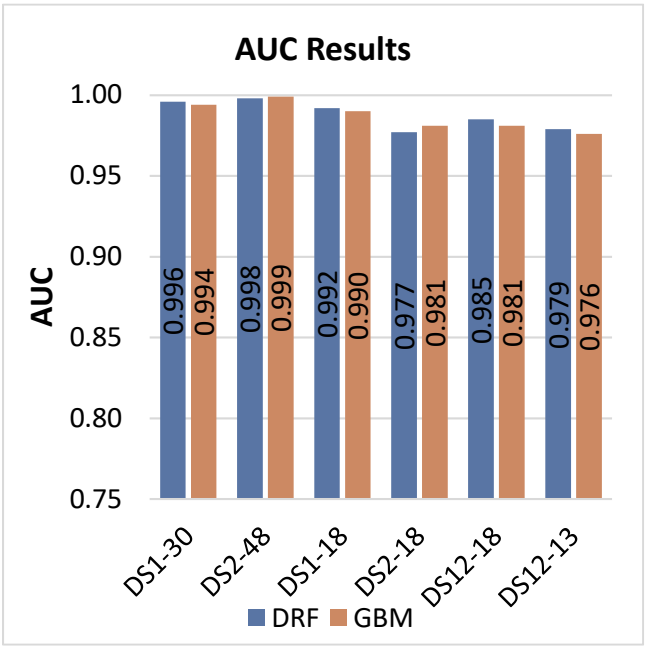
- Implementation
  - Best model built with 13 features was integrated into a Python application that takes a URLs as input, extracts features from the live website, and predicts if it is legitimate or phishing



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
    - DS1-18: DRF accuracy of **0.952** → **0.022** decrease from baseline
    - DS2-18: DRF accuracy of **0.937** → **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
    - DS12-13: DRF accuracy of **0.937**, AUC of **0.979**





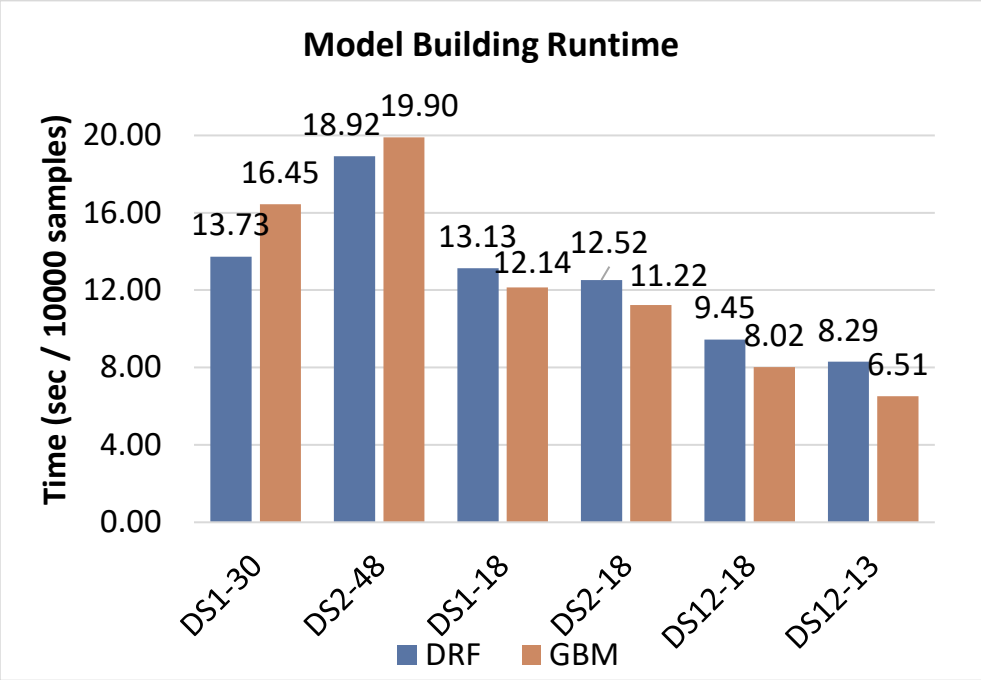
Results (2)

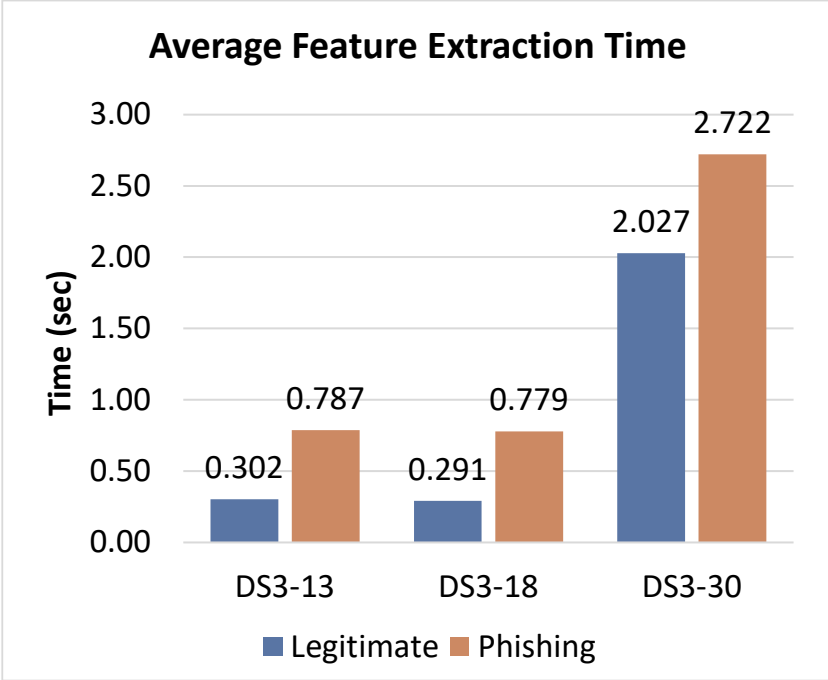
- Comparison with Previous Research
  - Second experiment compared the performance of DRF and GBM algorithms with best results achieved by previous research papers on the DS1-30 and DS2-48 datasets
  - Used the same validation techniques and data split ratio as reported by the authors of those papers
  - 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]	DS1-30	10f-CV	RF	**0.974**
Our results		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)

- 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!