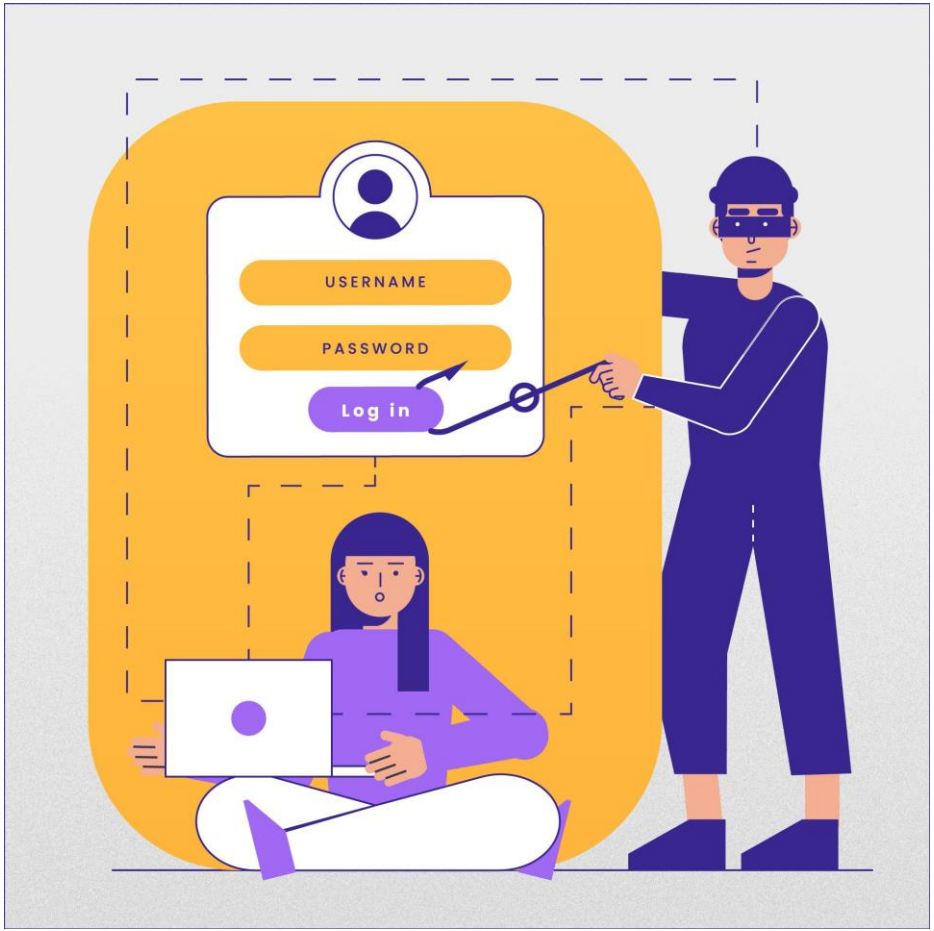


Catching Phishes with Machine Learning

Group 7:
Ang Su Yiin
Anne Nguyen Nhi Thai An
Gordy Adiprasetyo
Kendra Luisa Baylon Gadong

6 August 2021

Presentation Outline

- BACKGROUND AND MOTIVATION
- PROJECT OBJECTIVES
- PROJECT PIPELINE
- DATA PROCESSING
- MODEL SELECTION
- SUMMARY AND FUTURE WORK

Background and Motivation

https://dbs-transactionreview.com/

Phishing websites
Fake websites constructed to **look identical** to real sites with the intention of **tricking victims** into entering **login credentials**, which the phisher will then has access to.

Phishing cases on the rise

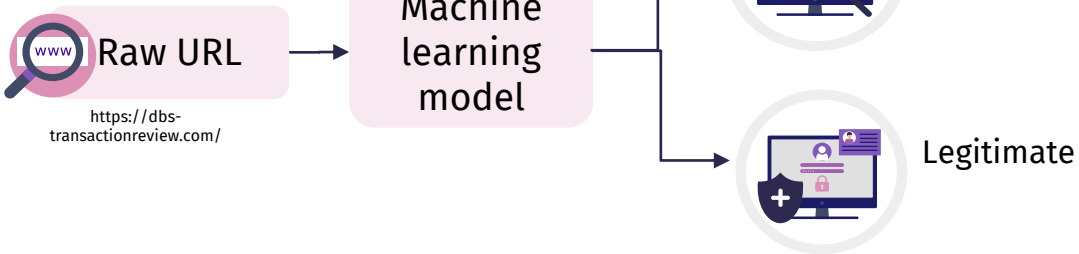
Number of **reported** phishing cases increased by **1240%** since 2016

2020 adjusted loss **> USD$54 mil**

Source: 2020 Internet Crime Report by FBI
https://www.ic3.gov/Media/PDF/AnnualReport/2020_IC3Report.pdf

Project Objectives

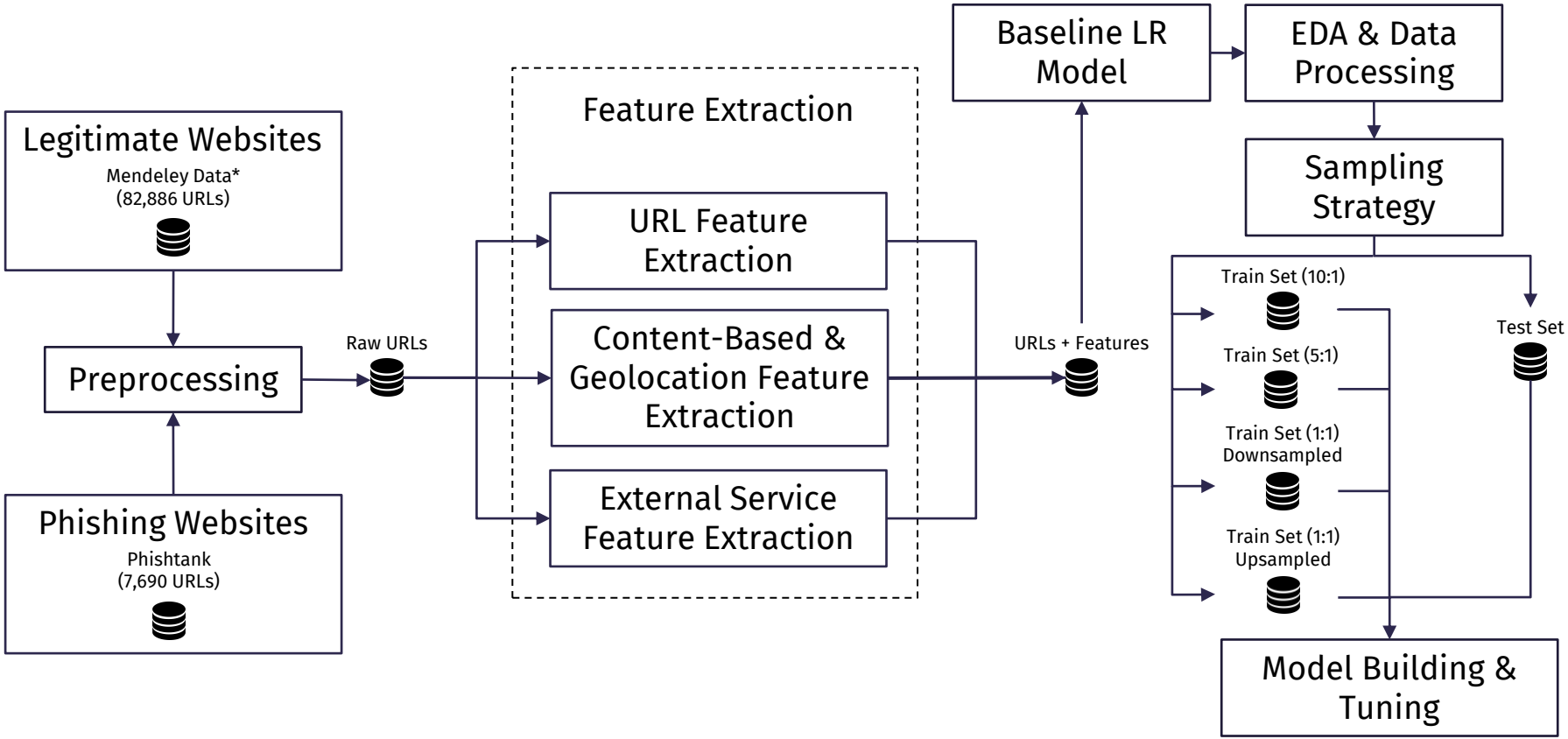
Detect **phishing websites** based on **raw URL** using machine learning techniques



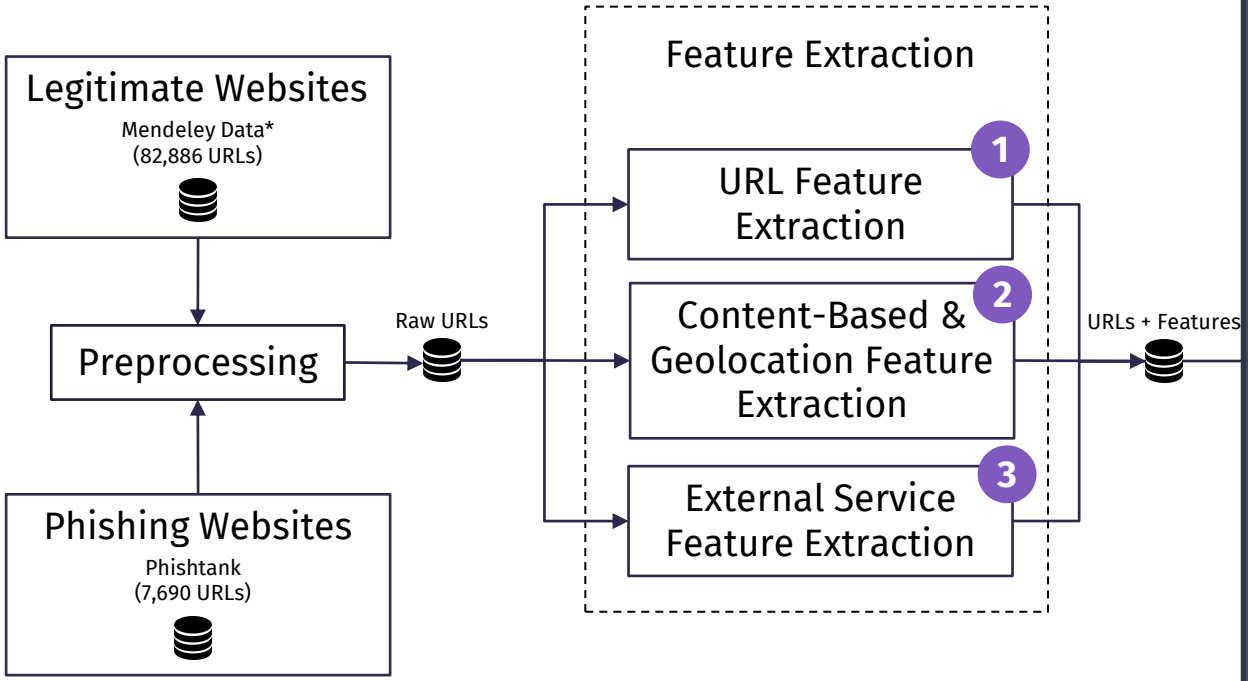
MAXIMIZE Recall

Model can be applied to :

- **Search engines**
- **Browser add-ons**
- Cybersecurity applications
- Mobile applications that link to external websites
- Platforms that allow embedding of external sites (forums, social media, etc)


Legitimate Websites
Mendeley Data*
(82,886 URLs)
Phishing Websites
Phishtank
(7,690 URLs)
Preprocessing
Raw URLs
Feature Extraction
URL Feature Extraction
Content-Based & Geolocation Feature Extraction
External Service Feature Extraction
URLs + Features
Baseline LR Model
EDA & Data Processing
Sampling Strategy
Train Set (10:1)
Train Set (5:1)
Train Set (1:1) Downsampled
Train Set (1:1) Upsampled
Test Set
Model Building & Tuning

Feature Extraction



1 URL Features

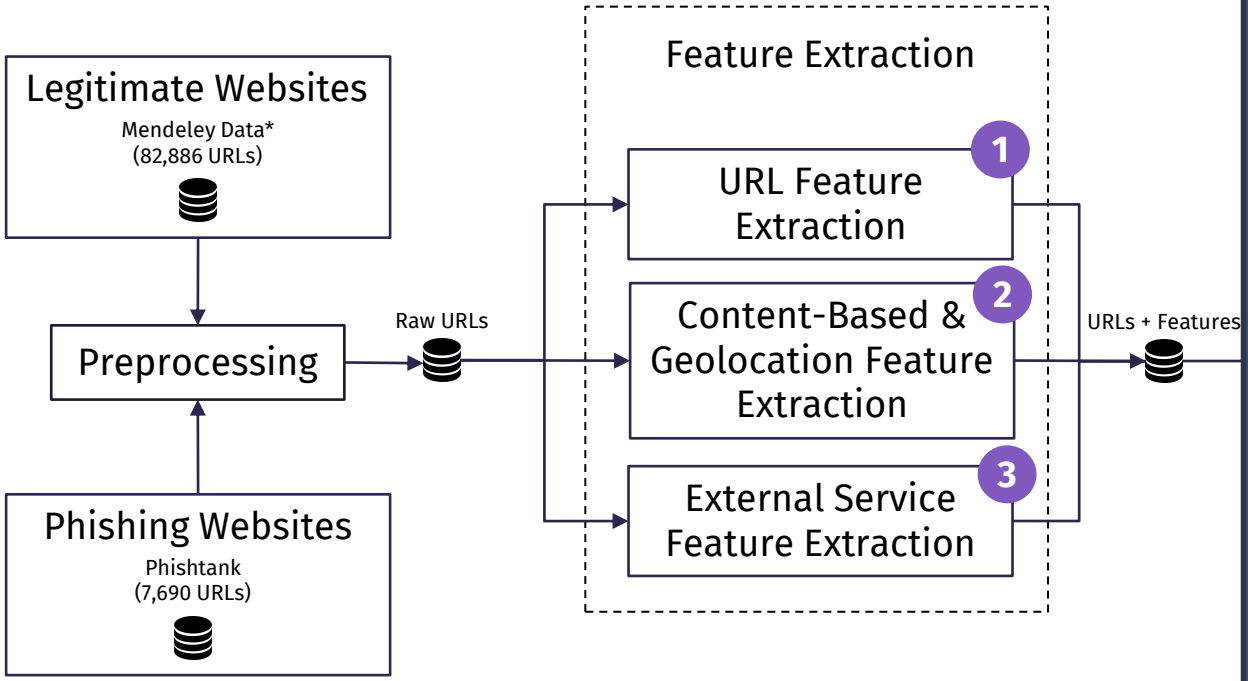
Features that are obtained by analyzing the text of URLs.

E.g. URL length, https usage, number of special characters, domains, usage of special extensions (e.g. .exe/.js)

Extracted a total of **55** URL features

https://data.mendeley.com/datasets/gdx3pkwp47/2
https://data.mendeley.com/datasets/c2gw7fy2j4/3

Feature Extraction



https://data.mendeley.com/datasets/gdx3pkwp47/2
https://data.mendeley.com/datasets/c2gw7fy2j4/3

2 Content-Based & Geolocation Features

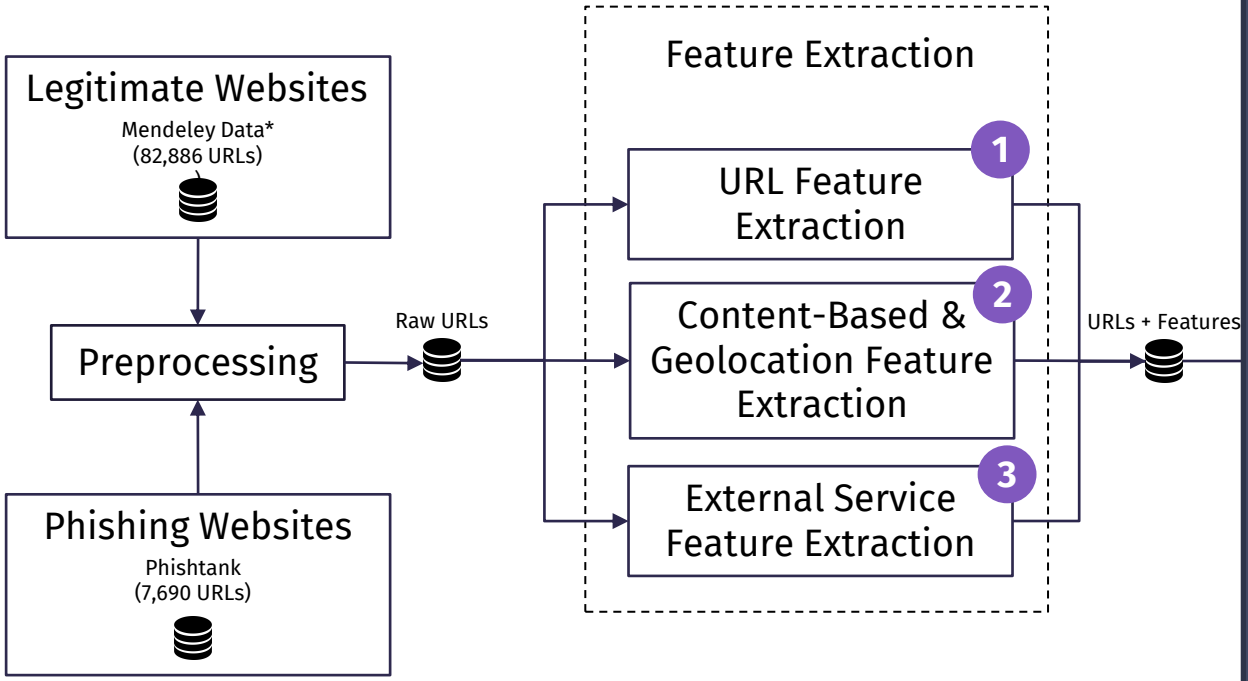
Features that are obtained by loading up the websites and analyzing their HTML contents

E.g. # of hyperlinks, presence of login forms, presence of pop-up windows

Extracted a total of **22** Content-based features

Using the websites' IP addresses, we also extracted Geolocation of hosting server

Feature Extraction



3 External Service Features

Features that are obtained by querying reference third party services and search engines.

E.g. WHOIS, Google, Openpagerank

Extracted a total of **7** external service features

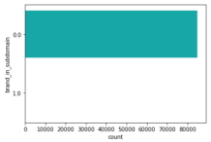
https://data.mendeley.com/datasets/gdx3pkwp47/2
https://data.mendeley.com/datasets/c2gw7fy2j4/3

EDA, Processing & Train-Test Split

Extracted data
90,576 rows x 89 columns:
- 3 id columns
- 1 target column
- 85 features

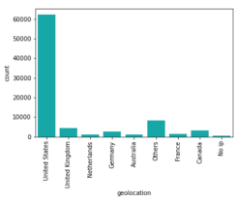
Remove 5,438 rows with **missing content features** data

Remove 7 features with **multi-collinearity**

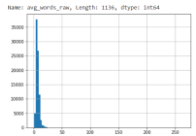


Remove 18 features with **single values/missing values**

Rebin/Recode



Standardize Numeric Features



One-hot Encode Categorical features

Processed data
85,138 rows x 68 features

Test Set (20%)

Train Set (80%)

Train Set (10:1)

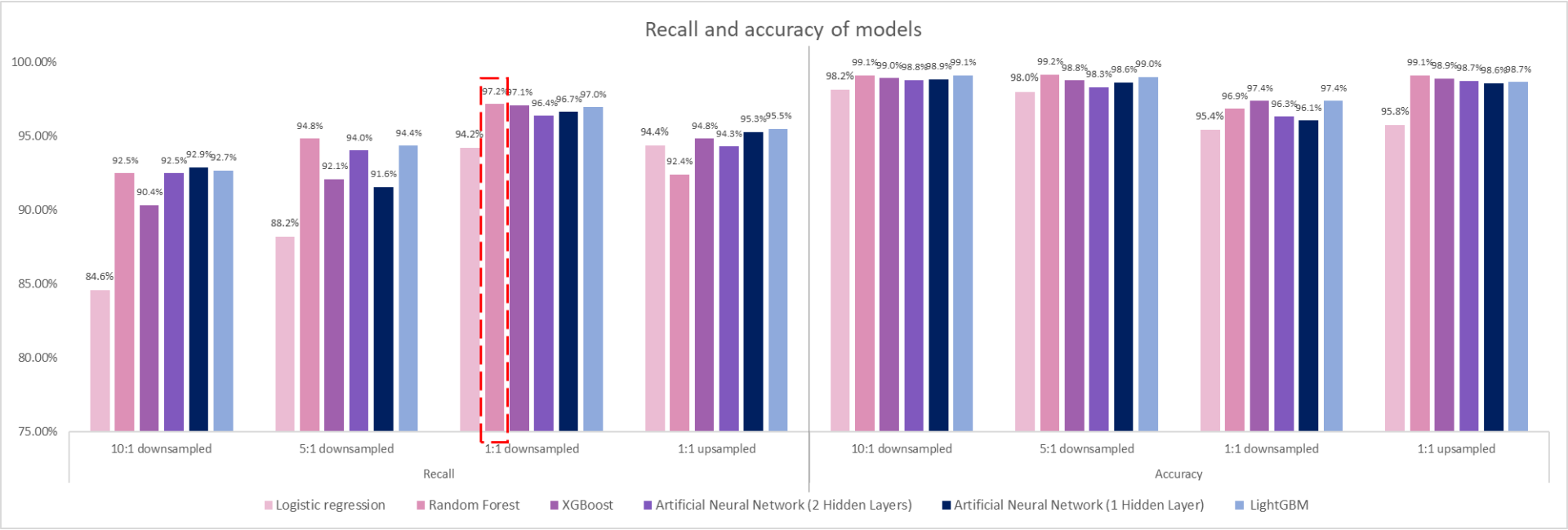
Train Set (5:1)

Train Set (1:1) Downsampled

Train Set (1:1) Upsampled

Model Evaluation and Selection

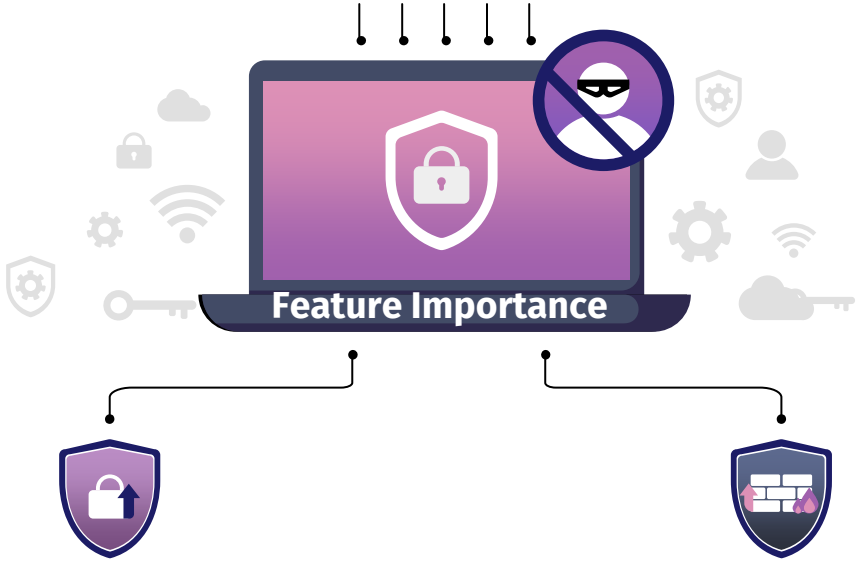
Random Forest (1:1 down-sampled) outperforms other models with recall score of 97.2%

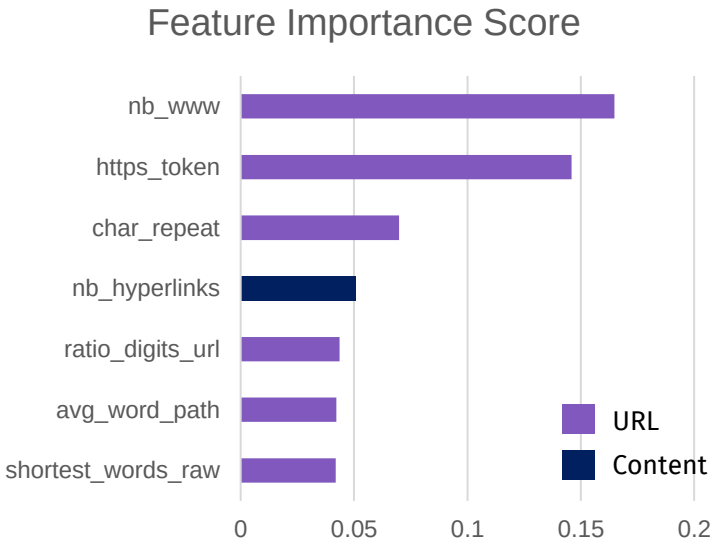


xx : 1 refers to training set ratio – xx legitimate : 1 phishing

Best Machine Learning Model

Random Forest (1:1 Legitimate : Phishing Ratio Down-sampled)





URL Features

nb_www | https_token
char_repeat | ratio_digits_url
avg_word_path | shortest_words_raw

Content-based Features

nb_hyperlinks

Summary and Future Work

Machine learning models are effective in detecting phishing websites due to characteristics of phishing websites.

Improve model

Collect more data on external features, update certain features e.g. phish hints

Improve feature selection - Balance cost of feature collection vs recall penalty

Explore model stacking

Deploy model

Browser plug-in for deployment

Thank you!
Questions?

Catching Phishes with Machine Learning

Group 7:
Ang Su Yiin
Anne Nguyen Nhi Thai An
Gordy Adiprasetyo
Kendra Luisa Baylon Gadong

Appendix I

Feature Description

URL-based Features

Feature	Description
length_hostname	hostname length
ip	IP address
nb_dots	no. of "."
nb_hyphens	no. of "-"
nb_at	no. of "@"
nb_qm	no. of "?"
nb_and	no. of "&"
nb_or	no. of "|"
nb_underscore	no. of "_"
nb_tilde	no. of "~"
nb_percent	no. of "%"
nb_slash	no. of "/"
nb_star	no. of "*"
nb_colon	no. of ":"
nb_comma	no. of ","
nb_dollar	no. of "$"
nb_space	no. of %20" or " "
nb_www	no. of "www"
nb_com	no. of ".com"
nb_dslash	no. of "//"
http_in_path	no. of "http"
https_token	use of https
ratio_digits_url	ratio of digits in full URL
ratio_digits_host	ratio of digits in hostname

Feature	Description
punycode	presence of a punnycode
port	port indicator
tld_in_path	top-level domain in path
tld_in_subdomain	top-level domain in subdomain
abnormal_subdomain	URLs matching a pattern of w[w]?[0-9]*
nb_subdomains	no. of subdomains
prefix_suffix	presence of "-" in domain names
shortening_service	use of a shortening service
path_extension	presence of a path extension
nb_redirection	number of redirections
nb_external_redirection	number of external redirections
length_words_raw	number of words
char_repeat	repeated characters
shortest_words_raw	shortest words in the url
shortest_word_path	shortest words in the path
longest_words_raw	longest words in the url
avg_word_host	average words in the hostname
avg_word_path	average words in the path
phish_hints	total occurrence of the ff phish hints: ‘wp’, ‘login’, ‘includes’, ‘admin’, ‘content’, ‘site’, ‘images’, ‘js’, ‘alibaba’, ‘css’, ‘myaccount’, ‘dropbox’, ‘themes’, ‘plugins’, ‘signin’, ‘view’
domain_in_brand	presence of brand in domain

Appendix I

Feature Description

Content-based and Geolocation Features

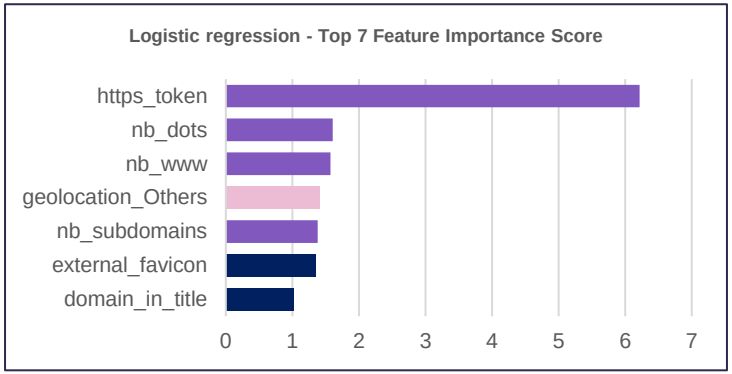
Feature	Description
nb_hyperlinks	no. of hyperlinks
ratio_intHyperlinks	ratio of internal hyperlinks
ratio_extHyperlinks	ratio of external hyperlinks
nb_extCSS	no. of external CSS files
login_form	presence of “”, “#”, “#nothing”, “#doesnotexist”, “#null”, “#void”, “#whatever”, “#content”, “javascript::void(0)”, “javascript::void(0);”, “javascript:::”, “javascript”
external_favicon	presence of external favicons
links_in_tags	ratio of internal links in <Link> tags
ratio_intMedia	ratio of internal media file links
ratio_extMedia	ratio of external media file links
safe_anchor	presence of ‘#’, ‘javascript’, or ‘mailto’ tags
empty_title	absence of web page title
domain_in_title	presence of domain of URL as part of web page title
domain_with_copyright	presence of domain of URLs within copyright logo
geolocation	country geolocation, based on IP address

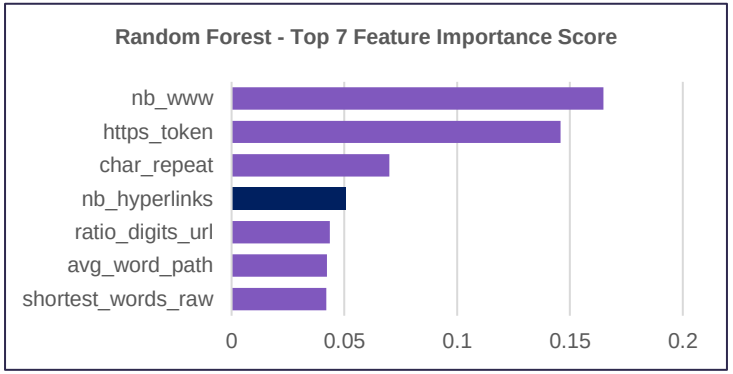
External Service Features

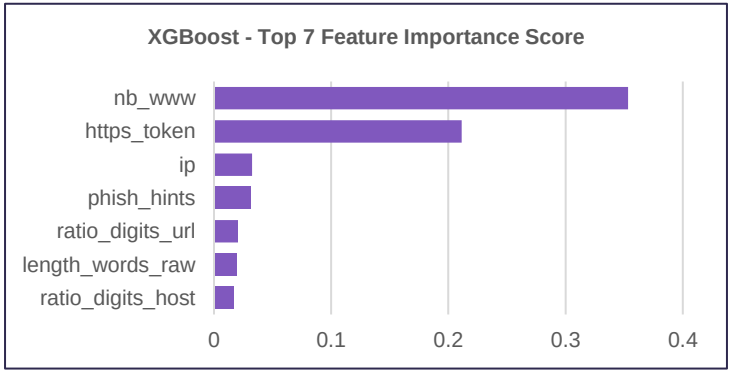
Feature	Description
web_traffic	number of visitors, retrieved from Alexa
dns_record	whether URL domain is registered within the DNS

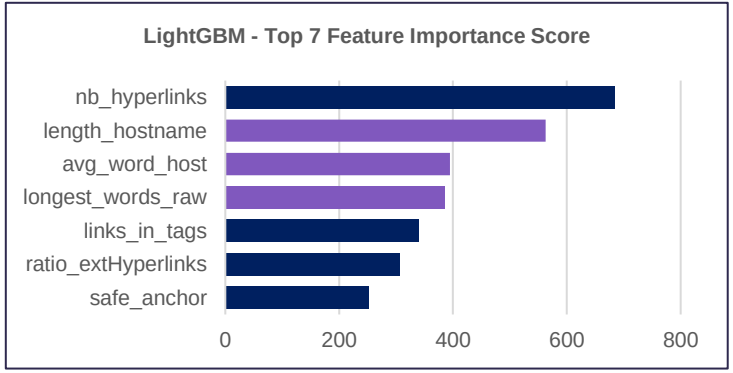
Appendix II

Feature Importance by Model









Feature type

- URL
- Content
- Geolocation

Unable to generate feature importance for ANN

Appendix III

Performance from literature

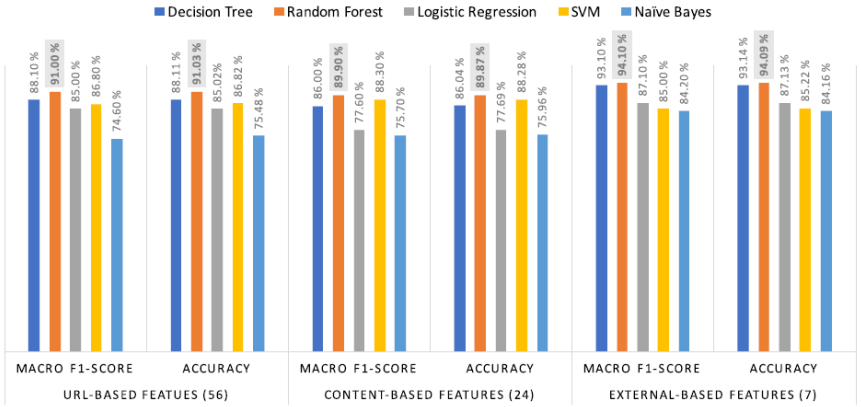


Fig. 8. Performance of classifiers trained on individual class of features.

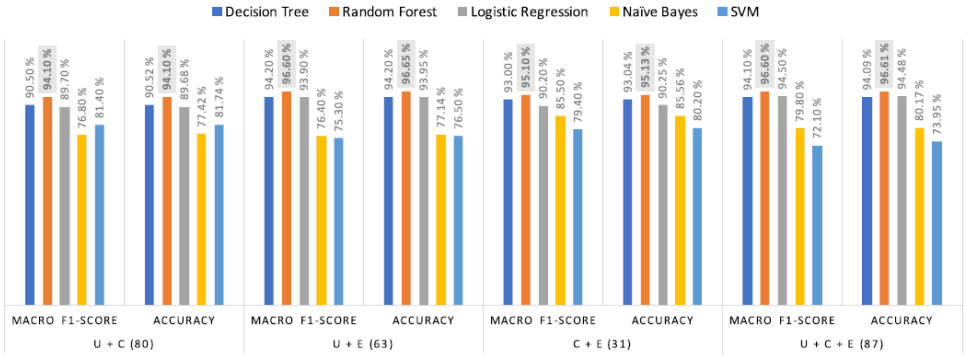


Fig. 9. Performance of classifiers trained on pairwise combined class of features.

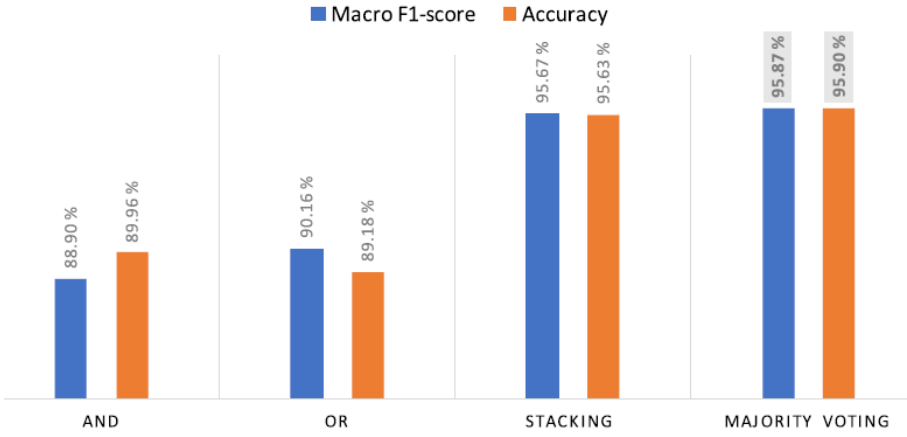


Fig. 10. Performance of combined models.

Hannousse, A., & Yahiouche, S. (2021). Towards benchmark datasets for machine learning based website phishing detection: An experimental study. *Engineering Applications of Artificial Intelligence*, *104*, 104347.