From Bytes to Bait:

Data Science in Phishing Detection

Emily Gelchie and Colin Fitzgibbons

Understanding Phishing

What is Phishing?

- ↦ Phishing is when attackers trick people into giving away personal or security information through fake emails or websites that look real.

Dangers of Phishing Sites:

- ↦ Steal Information
- ↦ Financial Loss
- ↦ Install Viruses
- ↦ Trust Issues

How to Stay Safe:

- ↦ Always check who's sending an email or message.
- ↦ Avoid clicking on suspicious links.
- ↦ Keep your security software updated.

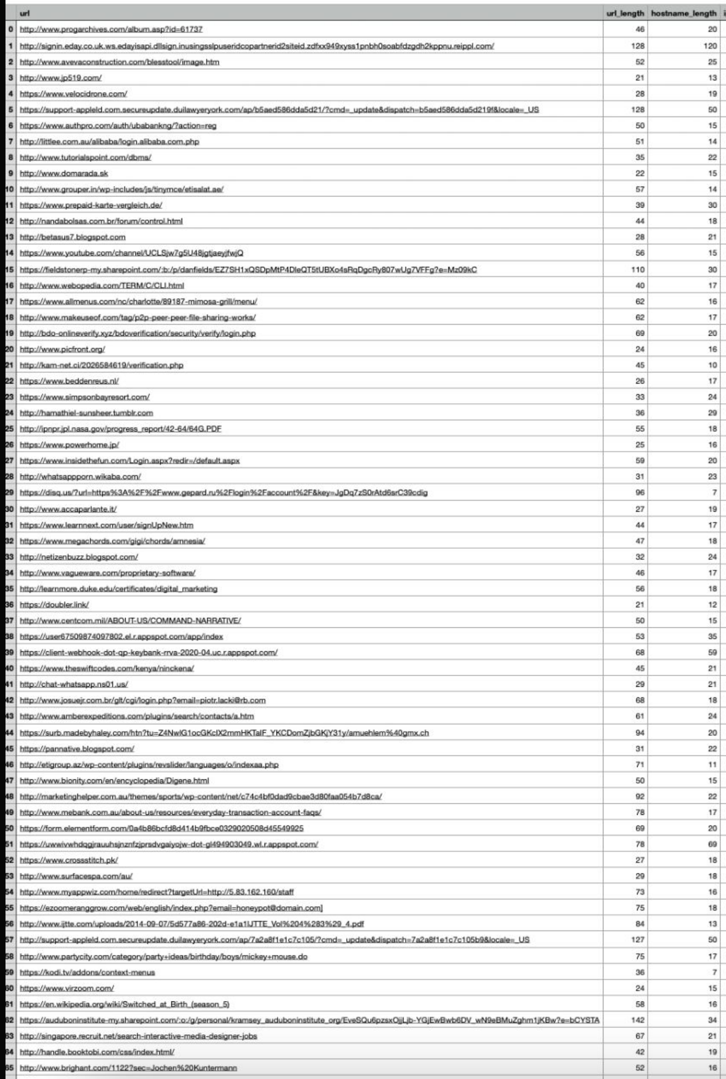
298,878

Individuals reported encountering Phishing Attacks in the United States alone during 2023 (Statista).

How can we build a machine learning model to detect phishing?

Our Datasets

1. **Dataset: dataset_link_phishing.csv**
   1. This dataset is a combination of two smaller datasets. It was used to train the machine learning model, ensuring a robust and comprehensive dataset for initial model learning.
   2. **Purpose:** To train the machine learning model to identify phishing sites
2. **Dataset: clean_newphishingdata.csv**
   1. This dataset serves as a new, independent test set. It was used to validate the machine learning model's accuracy and to check for overfitting.
   2. **Purpose:** To confirm the model's accuracy and generalizability by testing it against fresh, unseen data.



	url	url_length	hostname_length
0	http://www.progarchives.com/album.asp?id=61737	46	20
1	http://signin.eday.co.uk.ws.edayisapi.dllsign.inusingsslpuseridcopartnerid2siteid.zdfxx949xyss1pnbh0soabfdzgdh2kppnu.reippl.com/	128	120
2	http://www.avevaconstruction.com/blesstool/image.htm	52	25
3	http://www.jp519.com/	21	13
4	https://www.velocidrone.com/	28	19
5	https://support-appleld.com.secureupdate.duilawyeryork.com/ap/b5aed586dda5d21/?cmd=_update&dispatch=b5aed586dda5d219f&locale=_US	128	50
6	https://www.authpro.com/auth/ubabankng/?action=reg	50	15
7	http://littlee.com.au/alibaba/login.alibaba.com.php	51	14
8	http://www.tutorialspoint.com/dbms/	35	22
9	http://www.domarada.sk	22	15
10	http://www.grouper.in/wp-includes/js/tinymce/etisalat.ae/	57	14
11	https://www.prepaid-karte-vergleich.de/	39	30
12	http://nandabolsas.com.br/forum/control.html	44	18
13	http://betasus7.blogspot.com	28	21
14	https://www.youtube.com/channel/UCLSjw7g5U48jgtjaeyjfwjQ	56	15
15	https://fieldstonerp-my.sharepoint.com/:b:/p/danfields/EZ7SH1xQSDpMtP4DleQT5tUBXo4sRqDgcRy807wUg7VFFg?e=Mz09kC	110	30
16	http://www.webopedia.com/TERM/C/CLI.html	40	17
17	https://www.allmenus.com/nc/charlotte/89187-mimosa-grill/menu/	62	16
18	http://www.makeuseof.com/tag/p2p-peer-peer-file-sharing-works/	62	17
19	http://bdo-onlineverify.xyz/bdoverification/security/verify/login.php	69	20
20	http://www.picfront.org/	24	16
21	http://kam-net.ci/2026584619/verification.php	45	10
22	https://www.beddenreus.nl/	26	17
23	https://www.simpsonbayresort.com/	33	24
24	http://hamathiel-sunsheer.tumblr.com	36	29
25	http://ipnpr.jpl.nasa.gov/progress_report/42-64/64G.PDF	55	18
26	https://www.powerhome.jp/	25	16
27	https://www.insidethefun.com/Login.aspx?redir=/default.aspx	59	20
28	http://whatsappporn.wikaba.com/	31	23
29	https://disq.us/?url=https%3A%2F%2Fwww.gepard.ru%2Flogin%2Faccount%2F&key=JgDq7zS0rAtd6srC39cdig	96	7
30	http://www.accaparlante.it/	27	19
31	https://www.learnnext.com/user/signUpNew.htm	44	17
32	https://www.megachords.com/gigi/chords/amnesia/	47	18
33	http://netizenbuzz.blogspot.com/	32	24
34	http://www.vagueware.com/proprietary-software/	46	17
35	http://learnmore.duke.edu/certificates/digital_marketing	56	18
36	https://doubler.link/	21	12
37	http://www.centcom.mil/ABOUT-US/COMMAND-NARRATIVE/	50	15
38	https://user67509874097802.el.r.appspot.com/app/index	53	35
39	https://client-webhook-dot-qp-keybank-rrva-2020-04.uc.r.appspot.com/	68	59
40	https://www.theswiftcodes.com/kenya/ninckena/	45	21
41	http://chat-whatsapp.ns01.us/	29	21
42	http://www.josuejr.com.br/glt/cgi/login.php?email=piotr.lacki@rb.com	68	18
43	http://www.amberexpeditions.com/plugins/search/contacts/a.htm	61	24
44	https://surb.madebyhaley.com/htn?tu=Z4NwlG1ocGKclX2mmHKTalF_YKCDomZjbGKjY31y/amuehlem%40gmx.ch	94	20
45	https://pannative.blogspot.com/	31	22
46	http://etigroup.az/wp-content/plugins/revslider/languages/o/indexaa.php	71	11
47	http://www.bionity.com/en/encyclopedia/Digene.html	50	15
48	http://marketinghelper.com.au/themes/sports/wp-content/net/c74c4bf0dad9cbae3d80faa054b7d8ca/	92	22
49	http://www.mebank.com.au/about-us/resources/everyday-transaction-account-faqs/	78	17
50	https://form.elementform.com/0a4b86bcfd8d414b9fbce0329020508d45549925	69	20
51	https://uwwivwhdqgjrauuhsjnznfzjprsdvgaiyojw-dot-gl494903049.wl.r.appspot.com/	78	69
52	https://www.crossstitch.pk/	27	18
53	http://www.surfacespa.com/au/	29	18
54	http://www.myappwiz.com/home/redirect?targetUrl=http://5.83.162.160/staff	73	16
55	https://ezoomeranggrow.com/web/english/index.php?email=honeypot@domain.com]	75	18
56	http://www.ijtte.com/uploads/2014-09-07/5d577a86-202d-e1a1IJTTE_Vol%204%283%29_4.pdf	84	13
57	http://support-appleld.com.secureupdate.duilawyeryork.com/ap/7a2a8f1e1c7c105/?cmd=_update&dispatch=7a2a8f1e1c7c105b9&locale=_US	127	50
58	http://www.partycity.com/category/party+ideas/birthday/boys/mickey+mouse.do	75	17
59	https://kodi.tv/addons/context-menus	36	7
60	https://www.virzoom.com/	24	15
61	https://en.wikipedia.org/wiki/Switched_at_Birth_(season_5)	58	16
62	https://auduboninstitute-my.sharepoint.com/:o:/g/personal/kramsey_auduboninstitute_org/EveSQu6pzsxOjjLjb-YGjEwBwb6DV_wN9eBMuZghm1jKBw?e=bCYSTA	142	34
63	http://singapore.recruit.net/search-interactive-media-designer-jobs	67	21
64	http://handle.booktobi.com/css/index.html/	42	19
65	http://www.brighant.com/1122?sec=Jochen%20Kuntermann	52	16

Data Cleaning and Preprocessing

1

Original Dataset

The original dataset was cleaned for use by filtering out missing values, and converted the status of "Phishing" or "Legitimate" to a binary numeric variable, as well as looking for outliers and inconsistencies within other numeric columns.

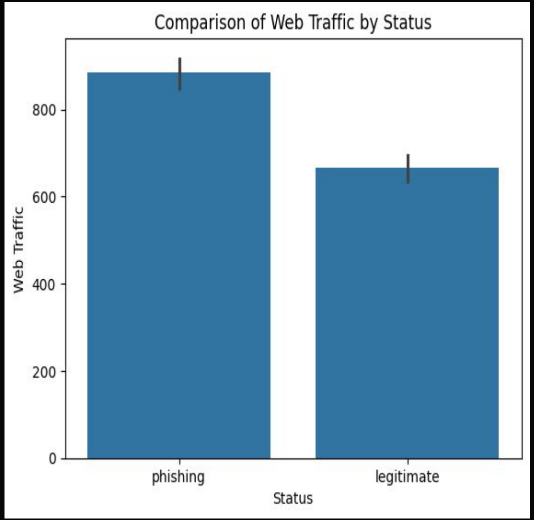
2

New Dataset

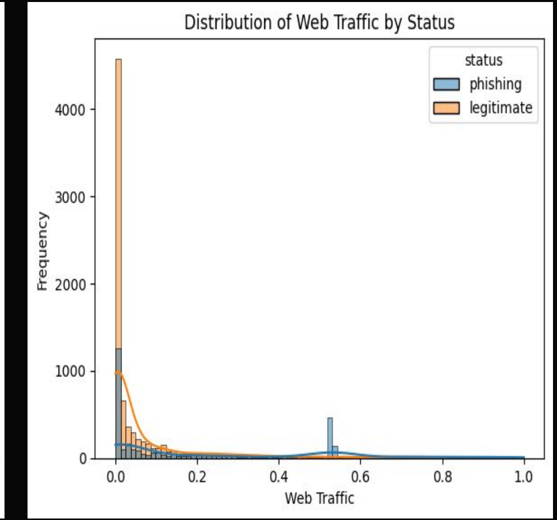
For the new dataset designed to test model accuracy and prediction, column names were changed to reflect that of the original dataset in regards to numeric variables. For example: "nb_colons" was changed to "total_of:"

Exploratory Data Analysis

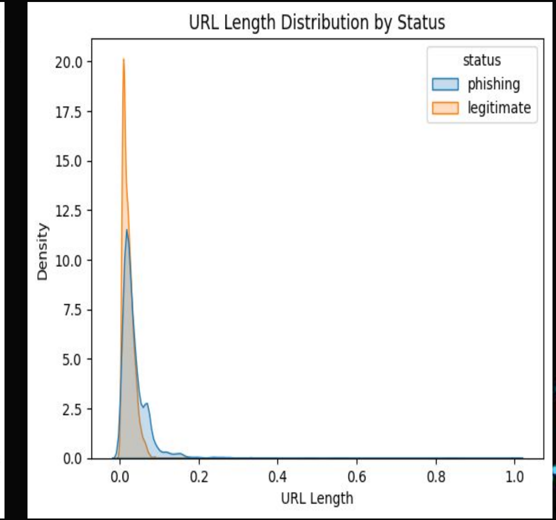
Web Traffic Comparison



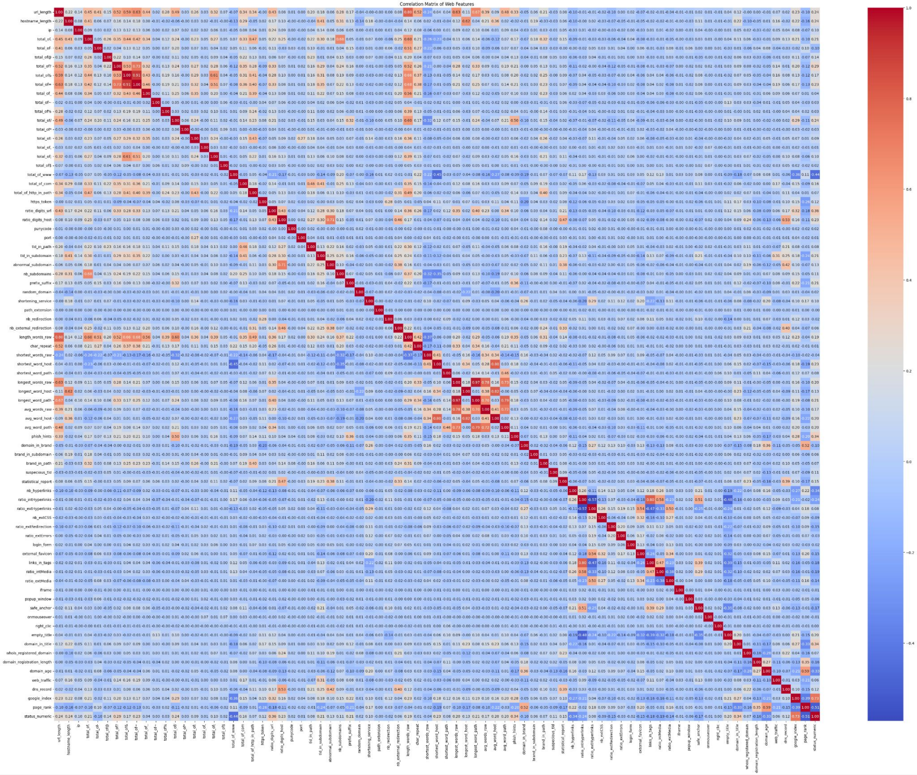
Web Traffic Distribution

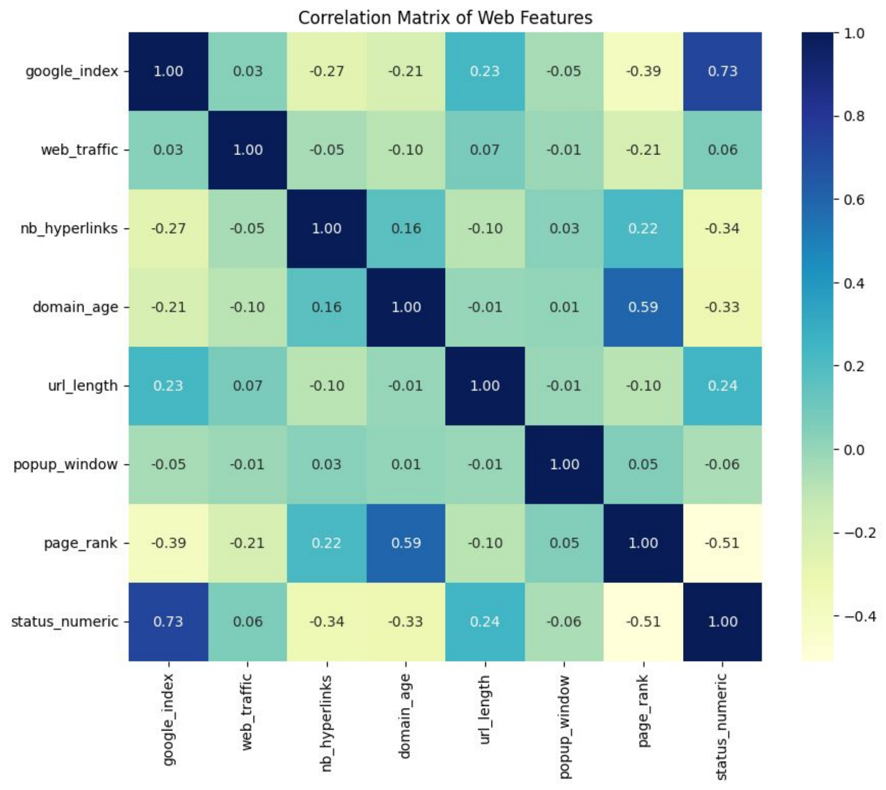


URL Length Distribution



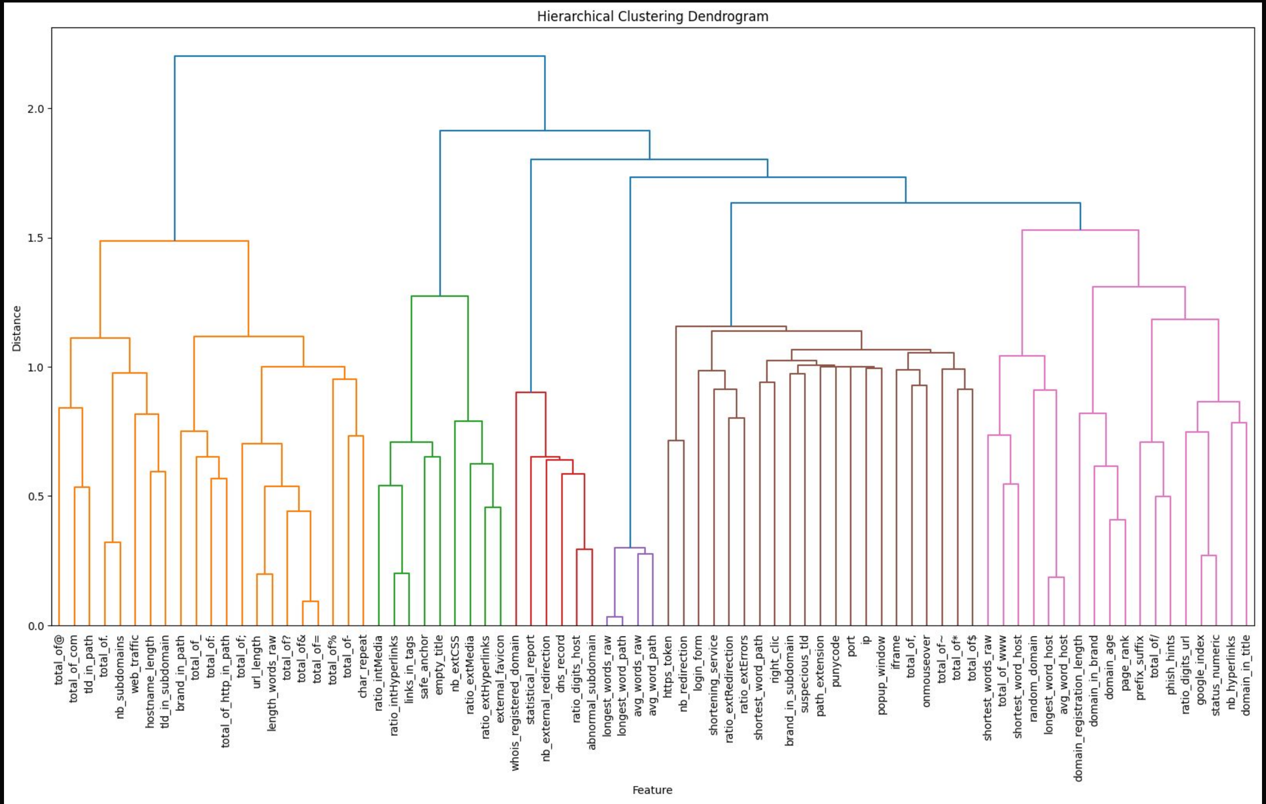
Exploratory Data Analysis





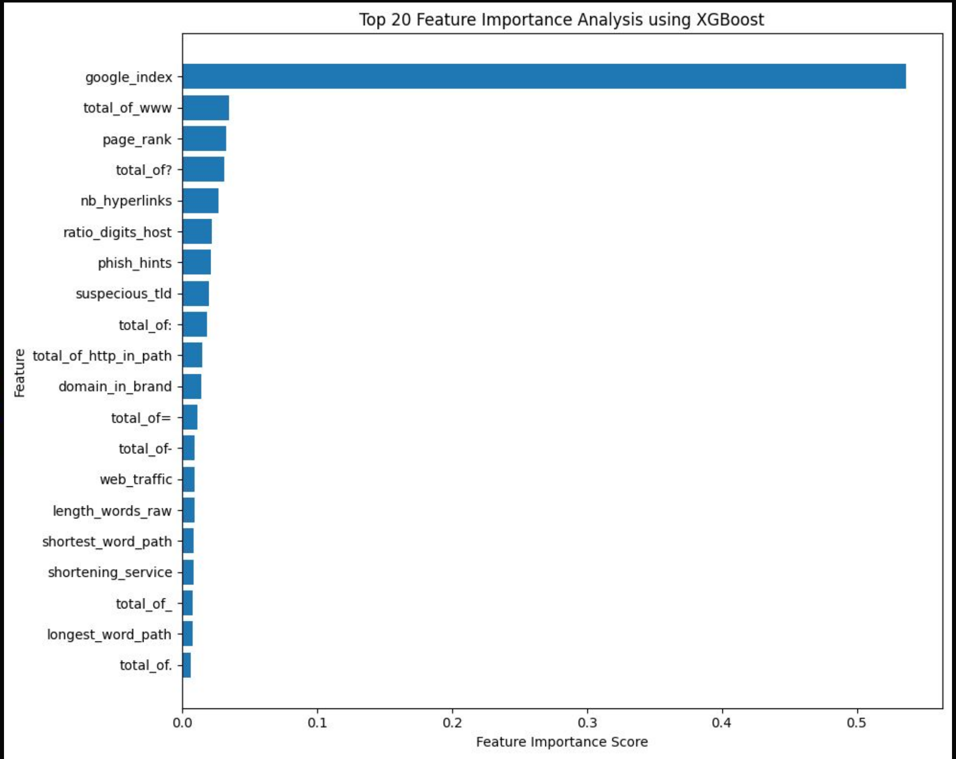
	google_index	web_traffic	nb_hyperlinks	domain_age	url_length	popup_window	page_rank	status_numeric
google_index	1.00	0.03	-0.27	-0.21	0.23	-0.05	-0.39	0.73
web_traffic	0.03	1.00	-0.05	-0.10	0.07	-0.01	-0.21	0.06
nb_hyperlinks	-0.27	-0.05	1.00	0.16	-0.10	0.03	0.22	-0.34
domain_age	-0.21	-0.10	0.16	1.00	-0.01	0.01	0.59	-0.33
url_length	0.23	0.07	-0.10	-0.01	1.00	-0.01	-0.10	0.24
popup_window	-0.05	-0.01	0.03	0.01	-0.01	1.00	0.05	-0.06
page_rank	-0.39	-0.21	0.22	0.59	-0.10	0.05	1.00	-0.51
status_numeric	0.73	0.06	-0.34	-0.33	0.24	-0.06	-0.51	1.00

Exploratory Data Analysis



Hierarchical Clustering Dendrogram

Feature Importance: XGBoost



Top Features:

google_index, total_of_www, page_rank, total_of?

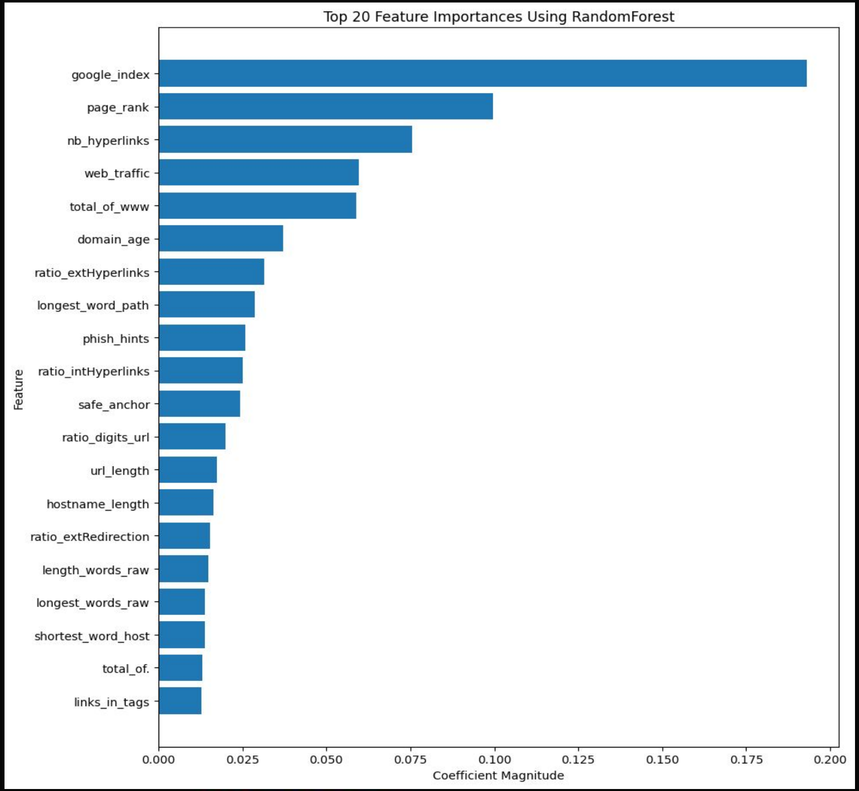
Feature Prioritization:

Gradient Boosting decision tree, "Gain" value

Significance:

XGBoost prioritizes google_index significantly over other variables within the dataset

Feature Importance: RandomForestClassifier



Top Features:

google_index, page_rank, nb_hyperlinks, web_traffic

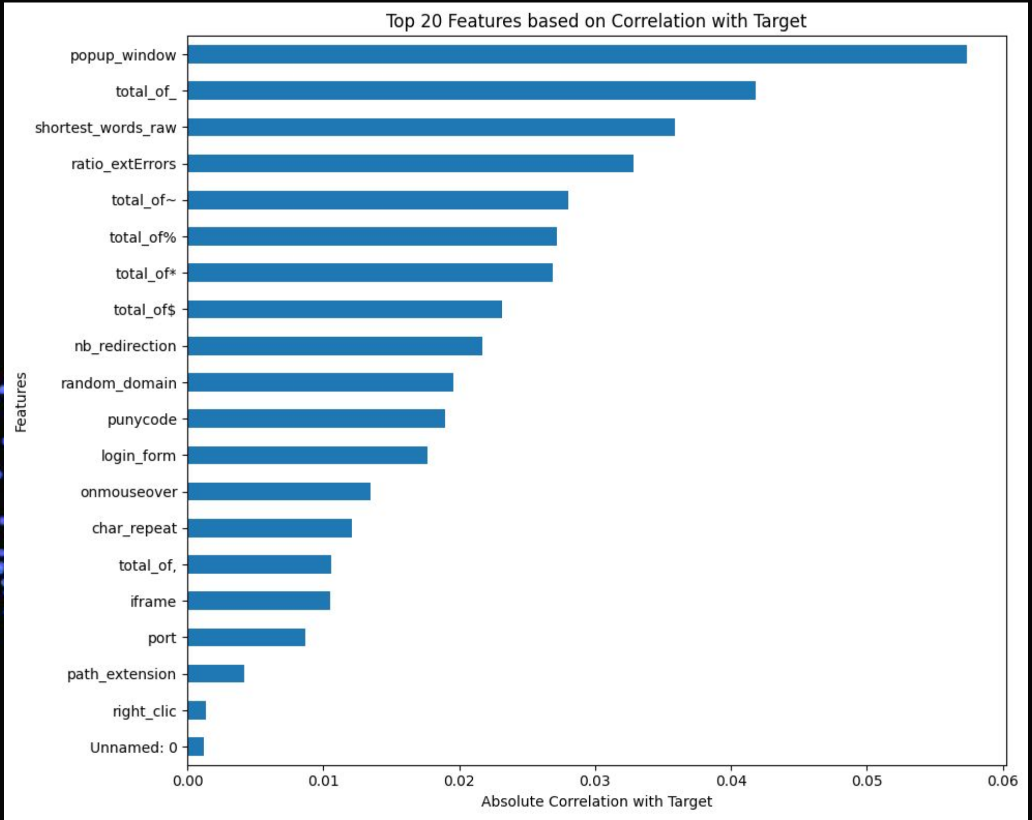
Feature Prioritization:

Randomly selects observations, builds a decision tree, and takes the average result

Significance:

RandomForestClassifier has more variables holding weight in the classification decision making process

Feature Importance: Correlation with Target



Top Features:

popup_window, total_of_, shortest_words_raw, ratio_extErrors

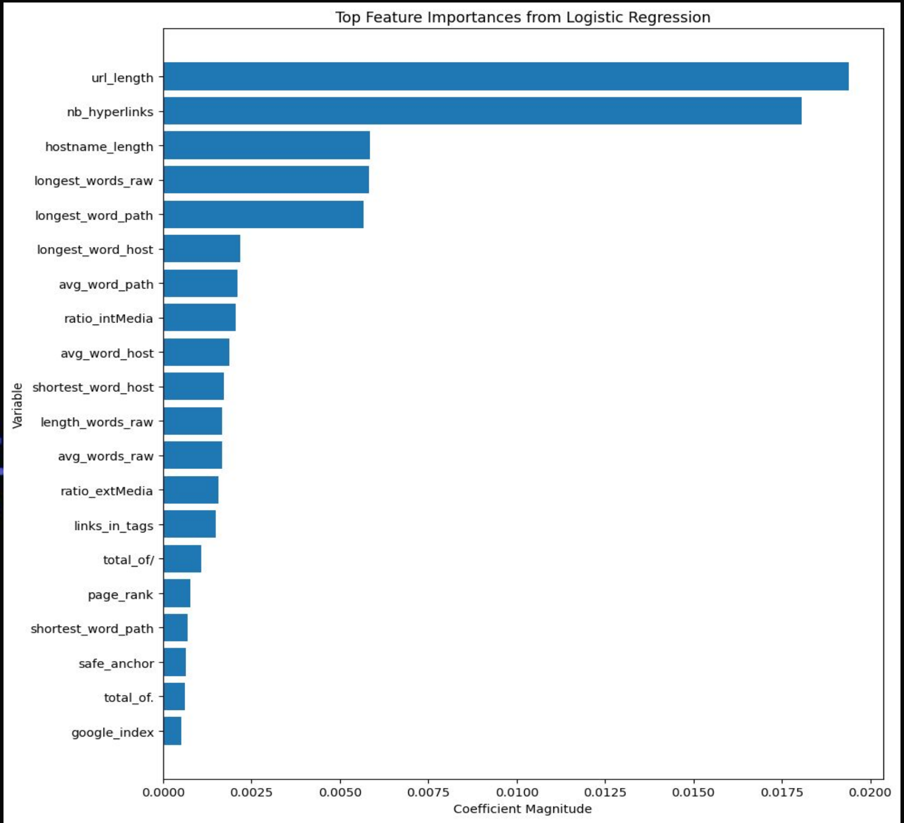
Feature Prioritization:

Computes correlation coefficients with target variable, and prioritizes features with higher absolute correlation coefficients.

Significance

Many features are used to make the decision, however, these features are different than the ones prioritized by other regression models

Feature Importance: Logistic Regression



Top Features:

url_length, nb_hyperlinks, hostname_length, longest_words_raw

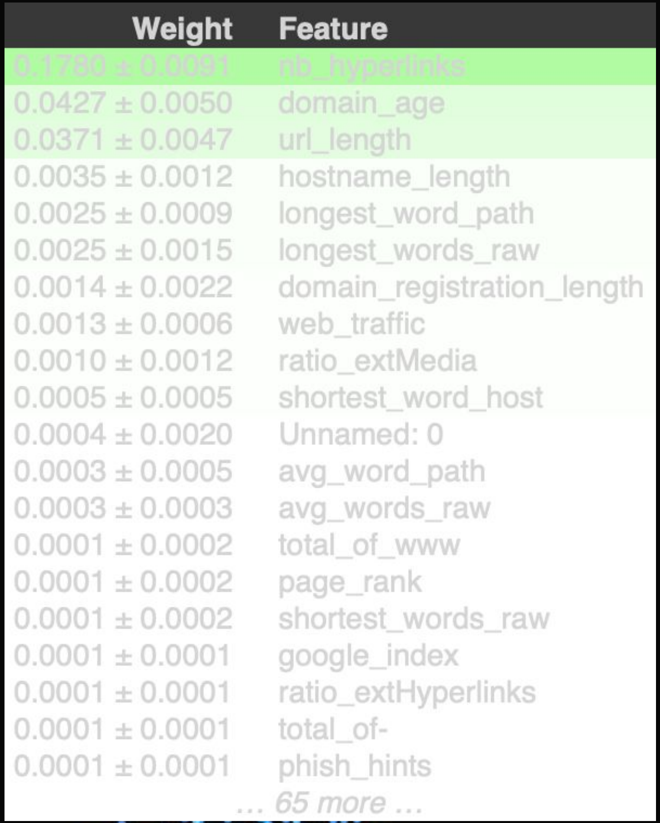
Feature Prioritization:

Examines coefficients and magnitudes based on model performance

Significance:

Logistic regression feature importance can provide a balance between interpretability, efficiency, and effectiveness.

Feature Importance: Permutation Importance



Weight	Feature
0.1780 ± 0.0091	nb_hyperlinks
0.0427 ± 0.0050	domain_age
0.0371 ± 0.0047	url_length
0.0035 ± 0.0012	hostname_length
0.0025 ± 0.0009	longest_word_path
0.0025 ± 0.0015	longest_words_raw
0.0014 ± 0.0022	domain_registration_length
0.0013 ± 0.0006	web_traffic
0.0010 ± 0.0012	ratio_extMedia
0.0005 ± 0.0005	shortest_word_host
0.0004 ± 0.0020	Unnamed: 0
0.0003 ± 0.0005	avg_word_path
0.0003 ± 0.0003	avg_words_raw
0.0001 ± 0.0002	total_of_www
0.0001 ± 0.0002	page_rank
0.0001 ± 0.0002	shortest_words_raw
0.0001 ± 0.0001	google_index
0.0001 ± 0.0001	ratio_extHyperlinks
0.0001 ± 0.0001	total_of-
0.0001 ± 0.0001	phish_hints

… 65 more …

Top Features:

nb_hyperlinks, domain_age, url_length, hostname_length

Feature Prioritization:

Weights projected by model reflect how much the performance would decline if the selected feature were removed.

Significance:

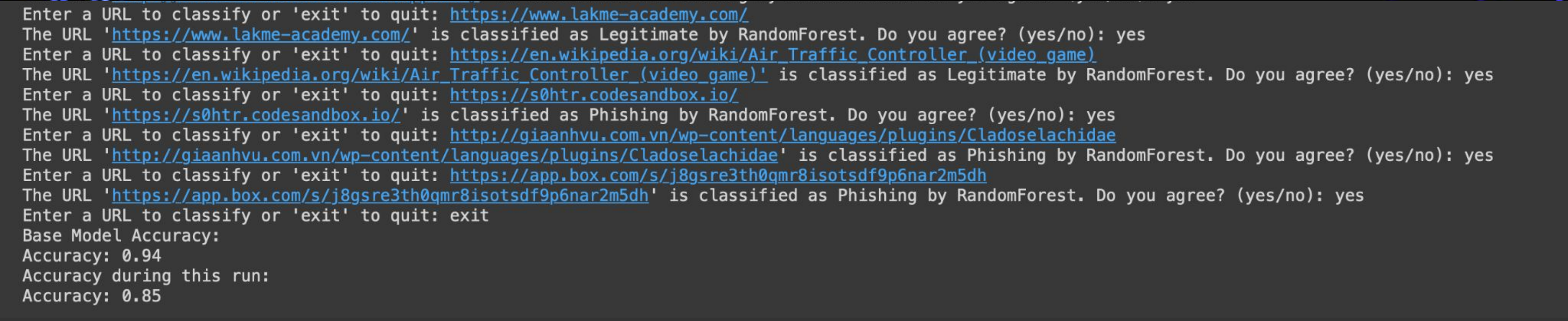
Useful for examining which features drive model performance and predictions and for identifying forms bias or overfitting.

Random Forest Machine Learning Model

- **Using RandomForestClassifier**
- We hand selected features we thought were the most important features based on the regression models
  - 'google_index', 'web_traffic', 'nb_hyperlinks', 'domain_age', 'url_length', 'popup_window', 'page_rank'
- **Accuracy for Original Dataset: 98.6% (0.9861075379470028)**
  - Root Mean Squared Error (RMSE): 0.11786628887428827
  - Mean Absolute Error (MAE): 0.01389246205299717
  - Precision: 0.9829
  - Recall: 0.9899
  - F1 Score: 0.9864
- **Accuracy for Secondary Dataset (Within same RandomForestClassifier): 99.7% (0.9979966901837819)**
  - RMSE: 0.04475
  - MAE: 0.002003
  - Precision: 0.997563
  - Recall: 0.998432
  - F1 Score: 0.997997

User Interactive Machine Learning

- User inputs a URL and the machine learning model tells the user if their link is phishing or legitimate
- Primary ML Model: RandomForestClassification
- Secondary ML Model (Reinforcement Learning): Stochastic Gradient Descent Classifier
- Features used:
  - 'url_length', 'hostname_length', 'https_token', 'nb_subdomains', 'prefix_suffix', 'tld_in_path', 'tld_in_subdomain', 'path_extension', 'random_domain', 'shortening_service', 'popup_window', 'total_of-', 'total_of@', 'total_of?', 'total_of&', 'total_of=', 'total_of_', 'total_of~', 'total_of%', 'total_of/', 'total_of*', 'total_of:', 'total_of,', 'total_of;', 'total_of$', 'total_of.', 'total_of_www'
- Base model is 94% accurate while each run the accuracy varies



```
Enter a URL to classify or 'exit' to quit: https://www.lakme-academy.com/
The URL 'https://www.lakme-academy.com/' is classified as Legitimate by RandomForest. Do you agree? (yes/no): yes
Enter a URL to classify or 'exit' to quit: https://en.wikipedia.org/wiki/Air_Traffic_Controller_(video_game)
The URL 'https://en.wikipedia.org/wiki/Air_Traffic_Controller_(video_game)' is classified as Legitimate by RandomForest. Do you agree? (yes/no): yes
Enter a URL to classify or 'exit' to quit: https://s0htr.codesandbox.io/
The URL 'https://s0htr.codesandbox.io/' is classified as Phishing by RandomForest. Do you agree? (yes/no): yes
Enter a URL to classify or 'exit' to quit: http://giaanhvu.com.vn/wp-content/languages/plugins/Cladoselachidae
The URL 'http://giaanhvu.com.vn/wp-content/languages/plugins/Cladoselachidae' is classified as Phishing by RandomForest. Do you agree? (yes/no): yes
Enter a URL to classify or 'exit' to quit: https://app.box.com/s/j8gsre3th0qmr8isotsdf9p6nar2m5dh
The URL 'https://app.box.com/s/j8gsre3th0qmr8isotsdf9p6nar2m5dh' is classified as Phishing by RandomForest. Do you agree? (yes/no): yes
Enter a URL to classify or 'exit' to quit: exit
Base Model Accuracy:
Accuracy: 0.94
Accuracy during this run:
Accuracy: 0.85
```

Challenges Faced

Cloud Computing

Our project involves multiple contributors, which prompted us to use Google Colab to facilitate collaborative access. However, this approach encountered challenges, including limitations on simultaneous code editing and recurring issues with document saving.

User Interactive Model Accuracy

The accuracy of our user-interactive model was compromised due to our inability to access all the data points used in training the machine learning model, as we lacked necessary API permissions.

Feature Engineering

Identifying crucial features across multiple regression models is challenging because each model applies its unique algorithmic approach to weigh the importance of features differently. This created a unique challenge when determining what to put in the final ML model.

Analyzing Binary Data

Due to the binary nature of the features in our categorical dataset, conducting exploratory data analysis and creating effective visualizations presented significant challenges. This limitation necessitated alternative approaches to data analysis to gain meaningful insights.

Next Steps

API Integration

Integrating an API into the user-centric machine learning model could significantly enhance accuracy by providing comprehensive access to all relevant variables from user-inputted sites, aligning closely with the training dataset.

Other Phishing

Phishing scams come in many different shapes and forms. A next step of the project could be to include LLMs to determine if text messages, phone calls, or emails are phishing or legitimate.

Expand Dataset

By integrating more data points into our model it will only increase the prediction accuracy.

Build an Application

Constructing a comprehensive database to differentiate phishing from legitimate sites, or by creating a user-friendly application that enables users to easily determine the legitimacy of websites, thereby enhancing the accessibility and utility of the code.

Key Project Takeaways

Analyzing binary and categorization data

Identifying and understanding feature importance

Building multiple regression models

Learning the strengths

Key Characteristics of phishing sites

Thank You

Questions?