JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Phishing detection using vector machine algorithm

Mayurdhvajsinh Raj, Smit Panchal, Nishant Khatri

Student of M.Sc. I.T, Student of MCA, Asst. Professor
Faculty of IT & Computer Science,
Parul University, Vadodara, India

Introduction

The web has become a platform to help various criminal companies like spam, financial fraud and operators spread malware. The correct commercial reasons for this plan may be different, but a common thread is a requirement that users don't have to visit their site. This visit should be possible using email, web query items, or connections from other site pages, however, the client must snap to make a move, for example, indicating the ideal URL (Uniform Resource Locator) and get significant data. To overcome this, the security community responded by developing a blacklist service packaged in toolbars, devices and search engines, providing warnings or alerts with accurate feedback. The site is too new, unclassified, or misclassified, so many harmful sites are not blacklisted.

Therefore, the phishing attack is a serious dangerous topic into cyber security field today because it allows attackers to obtain sensitive information from online users, such as personal identification numbers (PINs), MasterCard account information, login credentials. Phishing URLs primarily prey on individuals or groups of people using social engineering attacks that take advantage of people's lowered understanding of information security. These URLs entice internet users to visit fake websites so that their sensitive information, including debit and credit card data, can be collected. An attempt has been made to create a phishing detection method supported by lexical analysis of computer addresses and classifiers that use machine learning. On a dataset that originally had almost 1056937 labeled URLs, the tests were conducted (phishing and legitimate). The Support Vector Machine (SVM) classifiers, which had a rate of 99.89 percent accuracy in the police investigation of the examined URLs, outperformed Random Forest, Gradient Boosting, Neural Network, and other classifiers in the evaluation. We will create an online tool that uses artificial intelligence and machine learning to detect phishing websites. The objective is to grow into a platform with a sizable community that combats the expanding phishing attacks. This project's objectives include a single web application that enables users to identify phishing assaults. They offer their feedback by writing reviews and analyzing statistics concerning phishing attacks tested with it. The main contact type that gets sent to the admin, An admin panel to review the type of usage of the web app, Monitor users' reviews

Motivation

Phishing is increasing day by day , we have read few of the phishing attacks in history and in between 2013 and 2015, Facebook and Google were misdirected out of $100 million as a result of an extensive

phishing exertion. The phisher took advantage of the way that the two associations used Quanta, a Taiwan-based association, as a shipper. The assailant sent a movement of fake requesting to the association that imitated Quanta, which both Facebook and Google paid.

Eventually, the stunt was found, and Facebook and Google took action through the US general arrangement of regulations. The aggressor was caught and eliminated from Lithuania, and, as a result of the legal activities, Facebook and Google had the choice to recover $49.7 million of the $100 million taken from them.

Such type of incidents motivates us for developing such systems which can help user to avoid being getting phished.

Literature Review

The following list includes a few of the examples: The impact of AI on all these related domains will be seen in occupations that need imaginative and creative thinking, such as music and cuisine. With the use of AI, Cook Watson offers a glimpse into how the AI in the kitchen can take on the role of a sous chef to help create recipes and train their human counterparts. to create unique flavors. We merely hear 154, as we are prone to observe in daily life. SPAM. Some apps, like Gaana, create a personalized playlist of songs based on our prior listening habits. The usage of tools like Watson BEAT can provide composers with whole new musical aspects for inspiration. AI supports artists' perception[1]. The energy industry is a perfect example of how artificial intelligence (AI) and also machine learning (ML) are being used. It is obvious that industrial giants like BP and GE Power are abusing massive amounts of data and machine learning to improve performance and forecast operations for business optimization. Oil and gas production and refining efficiency, dependability, and safety are all improved by the use of technology. This illustrates how artificial intelligence can operate throughout the full energy range. Even while machine learning is still in its infancy, it has the potential to alter how we use our power. Both intelligent grids and the forecasting of renewable energy are affected by it. robots with intelligence (smart robots)[2]

The following examples, including robot-assisted surgery [3], virtual nursing assistants [4], and bodywork flow assistance [5], make it easy to understand how AI is used in this industry. – When it comes to AI-enabled applications with significant promise, robotic surgery is at the top of the list. By fusing historical operating statistics, real surgical expertise data, and pre-op case history information, AI-enabled artificial intelligence will enhance and direct the surgical instrument's accuracy [7].

In the last several years, numerous methods for identifying phishing assaults have been provided in the literature. We like to provide an overview of detecting strategies against phishing assaults in this area. User education-based strategies and software-based techniques are the two broad categories into which phishing detection methods may be broadly divided.

Heuristic-based, blacklist-based, and visual similarity-based approaches make up the three categories of software-based detection. Approaches focused on user education Phishing and non-phishing emails are categorized, [8]

created 2 embedded coaching modalities to demonstrate to consumers. Users will be able to identify phishing emails on their own with the help of this training. An educational interactive game called "Anti-Phishing Phill" was created by Sheng et al. [9] to teach smart practices to avoid phishing assaults. utilizing software-based methods The following sub-categories are further divided into for software-based detection: (a) Black-based approaches: with this kind of strategy, the suspect domain is compared to a blacklist of known phishing domains. The drawback of this theme is that not all phishing websites are always covered because it might take some time for a newly founded fraud website to be added to the blacklist. [10] showed

that blacklists are frequently added to records, with between 50 and 80 percent of phishing domains accessorial in blacklists when calculating some financial loss. (b) Heuristic-based approaches: In this type of strategy, the suspect website's heuristic design fits the feature set that is often present in phishing websites [11]. Attacks that are zero-day, or that have never been observed previously, are sometimes referred to as mistreatment heuristic techniques. [12] CANTINA, a content-based phishing detection method that uses an upscaled collection of characteristics from several fields on an online page, has been suggested. (c) Methods based on visual resemblance The visual appearance of a suspect website and its associated authentic website are compared using visual similarity-based techniques. Visual similarity-based approaches make decisions by considering sets of possibilities such as text content, markup language tags, cascading sheets (CSS), picture processing, etc. Bird genus and other terms [13] an extremely long web page with a proposed anti-phishing method backed by discriminative key-point characteristics. backed up the With the aforementioned methodologies proposed in the literature, we tend to realize that there isn't a single method that can identify different phishing attempts. Additionally, Blacklist/Whitelist-based techniques are not able to identify zero-day assaults. Heuristic-based solutions have a significant false positive rate and are only able to identify zero-day attacks. They also fail to identify attacks when an embedded item is given into the web page. Additionally, while visual similarity-based techniques may identify embedded items on a web page, they are unable to identify zero-day assaults. Therefore, in this article, we present PHISH-SAFE, a machine learning-based anti-phishing system that is supported by URL choices and may be able to quickly identify the kind of phishing assaults. Digital spam is frequently defined as "the commit to abuse, or manipulate, a techno social system by manufacturing and injecting unsolicited, and/or unwanted content geared toward influencing human behavior or the system itself, at the direct or indirect, immediate or long-term advantage of the spammer(s)."[8]. Spam mail is a general term for digital spam, but it may also be used to refer to any unsolicited, undesired, or trash email that comes from the receiver or any email that the user doesn't want to be included in their inbox. The analytic community has worked extremely hard over the past 20 years to reduce the spam mail problem, but the urgency has not diminished. Additionally, once spam is intended to deceive or influence on a large scale, it will alter society's structure and human behavior [14]. As a result, there has been a lot of spam mail recently, and according to [15], the amount has been growing every day in recent weeks. We looked at the potency of phishing and looked for remedies, focusing on two popular anti-phishing programs. They automated tests against a blacklist of 10,000 fictitious URLs maintained by Google and Microsoft for three weeks to evaluate the performance of the anti-phishing technologies built into Firefox 2 (i.e., Google blacklists) and Microsoft's Web Person 7. They also investigated the availability of page qualities that may be utilized to identify phishing pages by examining a huge range of phishing pages. and how these elements (links, dubious URLs, forms, and input fields) are used are frequently key factors in user fraud. [16]identifying specific and well-known detection holes by tracking the whole lifetime of significant phishing assaults. They created a cutting-edge architecture that enabled them to actively safeguard tens of thousands of accounts while passively monitoring victims' visits to phishing URLs. 4.8 million victims accessed phishing URLs during the course of a year, according to their network monitor, which does not count hunter trails. From the moment phishing campaigns joined to the network, through email distribution, Traveller tracking, scheme detection, and account engagement, they used these events and associated knowledge sources to analyze phishing efforts. They discovered that the average campaign lasts twenty-one hours from start to finish. Tourists present their credentials and ultimately their knowledge in a compromised and dishonest transaction at least 7.42% of the time. In addition, 89.13 percent of victims are the result of a tiny number of exceptionally no-hit campaigns. Six agreed with their conclusions. They draw attention to potential chances to counter those subtle strikes. [17] completed a scientific study of data from a large-scale real-world embedded phishing campaign using 115,080 phishing emails and 19,180 participants from one firm. The main goal of their investigation was to provide methods for addressing various biases in order to create a more reasonable evaluation of the efficacy of integrated phishing campaigns and coaching. The success of embedded phishing efforts is then examined using these techniques, and using the analysis, they look for ways to make those campaigns' planning more effective.

Using their approach incorrectly, they discovered that improvements in training appeared to be limited to more convincing phishing emails and that there was no improvement for less appealing phishing emails. backed up their results, Hey will suggest changes to the integrated phishing campaign's layout that can boost its strength and efficiency.[18] explored how well phishing training and education work in helping users recognize distinct phishing attacks. Users' abilities to spot bogus emails, SMS phishing (Smishing), dishonest phone calls (Vishing), and phishing via social networks were evaluated. The study's objectives were to evaluate online anti-phishing educational resources and users' capacity to recognize phishing attacks. To do this, a phishing form was created to perform a pre- and post-test experiment to see if there was a significant difference between the participants' average pre- and post-scores after receiving phishing education and training materials. No significant changes were seen in the test scores of the millions of forty-three individuals after participants received phishing instruction, according to the analysis findings. The study examined variables that could influence or have an impact on the findings, such as difficulties in comprehending phishing teaching materials. However, any study is necessary to address these issues, and a number of other research directions are being investigated. [19] To examine the effects of habitual hazardous behaviors that make a target more likely to run into a motivated offender, we employed an integrative mode exposure model. Knowledge gathered in 2016 from a sampling (n = 723) was used to these goals. And other factors that are quite important in victimization were examined. There is a connection between phishing, digital repeated behavior, and internet browsing. Additionally, a connection between impulsivity and every online activity (apart from online buying behavior) was discovered. According to this study, specific internet users and consumers who often exchange and utilize files found online need to be taught how to recognize and respond to phishing scams. Targeted phishing attacks affected over 90% of businesses in 2019, which is a significant cause for worry given that they are now a global problem. Concern over the matter increased as a result of a 67 percent increase in the amount of correspondence via email. Those emails' primary goal is to protect consumers from frauds because it has been observed that the vast majority of people don't have the cybersecurity knowledge necessary to recognize these scams. For instance, 90% of working individuals admitted to using company-issued computers for personal use, 50% of them admitted to not password-protecting their home networks, and 45% of them admitted to using arcanum reuse.[20]

Summary

User education-based strategies and software-based techniques are the two broad categories into which phishing detection methods may be broadly divided. utilizing software-based methods The following sub-categories are further divided into for software-based detection: (a) Black-based approaches: with this kind of strategy, the suspect domain is compared to a blacklist of known phishing domains. Additionally, Blacklist/Whitelist-based techniques are not able to identify zero-day assaults. Additionally, while visual similarity-based techniques may identify embedded items on a web page, they are unable to identify zero-day assaults. Therefore, in this article, we present PHISH-SAFE, a machine learning-based anti-phishing system that is supported by URL choices and may be able to quickly identify the kind of phishing assaults. The main goal of their investigation was to provide methods for addressing various biases in order to create a more reasonable evaluation of the efficacy of integrated phishing campaigns and coaching. backed up their results, They will suggest changes to the integrated phishing campaign's layout that can boost its strength and efficiency.[18] explored how well phishing training and education work in helping users recognize distinct phishing attacks. The study's objectives were to evaluate online anti-phishing educational resources and users' capacity to recognize phishing attacks.

Methodology

The ongoing method for phishing area techniques encounters low distinguishing proof accuracy and high deceptive issue especially when different phishing approaches are introduced. Above and beyond, the most generally perceived methodology used is the blacklist based procedure which is inefficient in noting transmitting phishing attacks since enrolling another space has become more direct, no expansive blacklist can ensure an ideal extraordinary informational collection for phishing area

The proposed phishing acknowledgment structure utilizes AI models and significant cerebrum associations. The structure includes two critical parts, which are the AI models and a web application. These models include Decision Tree, Support Vector Machine, XGBooster, Multilayer Perceptions, Auto Encoder Neural Network, and Random Forest. These models are picked after different assessment based presentations of various AI computations. All of these models is ready and taken a stab at a site content-based incorporate, removed from both phishing and genuine dataset. In this manner, the model with the most significant precision is picked and integrated into a web application that will enable a client to predict if a URL interface is phishing or genuine.

Benefits of the new structure

i. Will really need to isolate some place in the scope of phishing(0) and legitimate(1) URLs
ii. It Will help decline phishing data breaks for an affiliation
iii. It Will be valuable to individuals and affiliations
iv. It is easy to use

The model improvement procedure takes a couple of models, tests them, and adds them to an iterative collaboration until a model that meets the normal necessities is made. Coming up next are the stages to AI model improvement for phishing distinguishing proof structures:

1. Data Collection

The data used to make the datasets on which the models are arranged are gotten from different open-source stages. The dataset combination includes of phishing and legitimate URL dataset. The course of action of phishing URLs are accumulated from an open-source organization called Phish Tank. This help gives a lot of phishing URLs in various plans like CSV, JSON, and so forth that gets invigorated hourly. This dataset is accessible from the phishtank.com site. From this dataset, more than 5000 unpredictable phishing URLs are accumulated to set up the ML models The course of action of legitimate URLs are gotten from the open datasets of the School of New Brunswick, This dataset is accessible on the school site. This dataset has a combination of innocuous, spam, phishing, malware and deformation URLs. Out of this large number of types, the innocuous URL dataset is considered for this undertaking. From this dataset, Over 5000 inconsistent genuine URLs are accumulated to set up the ML models.

2. Preprocessing

Data preprocessing is the first and critical stage after data combination. The unrefined dataset obtained for phishing distinguishing proof was prepared by dispensing with monotonous additionally, inconsistent data and moreover encoded using the One-Hot Encoding methodology into a supportive and capable design sensible for the AI model.

3. Exploratory data assessment

Exploratory data assessment (EDA) technique was used on the dataset later series of data cleaning. The data discernment procedure was used to research, examine and summarize the dataset. These discernment involve heat-map, histograms, box plots, disseminate plots, and match plots to uncover models and pieces of information inside data.

Feature Extraction

Feature Extraction intends to decrease the number of components in a dataset by making new components from the ongoing ones. As such, Website content-based features were eliminated from phishing and certifiable datasets, for instance, the Address bar-based incorporate which includes 9 components, Domain-based incorporate which contains 4 features, and HTML and JavaScript-based Feature which includes 4 components. Along these lines, completely 17 features were isolated for phishing acknowledgment.

4. Model Training

Model Training incorporates dealing with Machine learning computations with data to help perceive and learn extraordinary credits of the dataset. This assessment issue is a consequence of overseen understanding, which falls under the portrayal issue. The computations used for phishing area include controlled AI models (4) and significant cerebrum association (2) which was used to set up the dataset. These computations consolidate Decision Tree, Random Forest, Support Vector Machines, XGBooster, Multilayer Perceptron, and Auto-encoder Neural Network. This enormous number of models were ready on the dataset. As such, the dataset is spitted into a readiness and testing set. The arrangement model includes 80% of the dataset to enable the machine learning models to all the more profoundly concentrate on the data and have the choice to perceive among phishing and certifiable URLs

5. Model Testing

Model Testing incorporates the cycle where the introduction of a totally ready model is evaluated on a testing set. In this manner, after 80% of data has been arranged, 20% of the dataset is used to evaluate the arranged dataset to see the presentation of the models.

6. Model Evaluation

Model Evaluation incorporates surveying the hypothesis accuracy of models and finishing up whether or not the model performs better. Subsequently, Scikit-learn (sklearn metrics) module was used to completes a couple of score and utility capacities to measure the request execution to fittingly evaluate the models conveyed for phishing revelation.

Datasets

This dataset contains 48 features removed from 5000 phishing site pages and 5000 certifiable site pages, which were downloaded from January to May 2015 and from May to June 2017. A predominant part extraction technique is used by using the program robotization framework (i.e., Selenium WebDriver), which is more definite and solid stood out from the parsing approach considering standard verbalizations.

Antagonistic to phishing researchers and experts could find this dataset significant for phishing features assessment, coordinating quick confirmation of thought tests or benchmarking phishing gathering models.

url	# length_url	# length_hostname	# ip	# nb_dots
11429 unique values	12 – 1641	4 – 214	0 – 1	1 – 24
http://www.crestonwood.com/router.php	37	19	0	3
http://shadetreetechnology.com/V4/validation/a111aedc8ae390eabcfa130e041a10a4	77	23	1	1
https://support-appleld.com.secureupdate.duilawyeryork.com/ap/89e6a3b4b063b8d/?cmd=_update&dispatch	126	50	1	4

Figure 1: dataset 2

The outfitted dataset consolidates 11430 URLs with 87 eliminated features. The dataset is expected to be used as benchmarks for AI based phishing revelation structures. Features are from three one of a kind classes: 56 eliminated from the development and accentuation of URLs, 24 isolated from the substance of their columnist pages, and 7 are removed by addressing external organizations. The dataset is changed, it contains unequivocally half phishing and half genuine URLs.

# Index	# UsingIP	# LongURL	# ShortURL	# Symbol@
It denotes the index number	(categorical - signed numeric) : { -1,1 }	(categorical - signed numeric) : { 1,0,-1 }	(categorical - signed numeric) : { 1,-1 }	(categorical - signed numeric) : { 1,-1 }
0 – 11.1k (10500.35 - 11053.00, Count: 553)	-1 – 1	-1 – 1	-1 – 1	-1 – 1
0	1	1	1	1
1	1	0	1	1
2	1	0	1	1
3	1	0	-1	1
4	-1	0	-1	1
5	1	0	-1	1
6	1	0	1	1

Figure 2: dataset 3

The data set is available both in text and csv files which provides the following resources that can be used as inputs for model building: A collection of website URLs for 11000+ websites. Each sample has 30 website parameters and a class label identifying it as a phishing website or not (1 or -1). The data set also serves as an input for project scoping and tries to specify the functional and non-functional requirements for it.

phish_id	url	phish_detail_url	submission_time	verified	verification_time	online	target
7646291	https://benjaminhope739.wixsite.com/my-site-1	http://www.phishtank.com/phish_detail.php?phish_id=7646291	2022-08-10T10:03:45+00:00	yes	2022-08-10T10:10:39+00:00	yes	Other
7646290	https://4attwebmaillahsydb34455357 8attwebmaillahsydb355646654.weebly.com/	http://www.phishtank.com/phish_detail.php?phish_id=7646290	2022-08-10T10:03:40+00:00	yes	2022-08-10T10:10:39+00:00	yes	Other
7646275	https://bt-service-104877.weeblysite.com/	http://www.phishtank.com/phish_detail.php?phish_id=7646275	2022-08-10T09:43:22+00:00	yes	2022-08-10T09:52:44+00:00	yes	Other
7646265	https://escuelarelonca vi.cl/nhcb.bns/5/nssb.php	http://www.phishtank.com/phish_detail.php?phish_id=7646265	2022-08-10T09:32:27+00:00	yes	2022-08-10T09:42:38+00:00	yes	Other
7646260	https://veoveo.com.vn/wp-content/backups-dup-lite/985023jtvsrsd/	http://www.phishtank.com/phish_detail.php?phish_id=7646260	2022-08-10T09:30:12+00:00	yes	2022-08-10T09:42:38+00:00	yes	Other
7646259	https://banking.dlkb.eu/	http://www.phishtank.com/phish_detail.php?phish_id=7646259	2022-08-10T09:30:08+00:00	yes	2022-08-10T09:42:38+00:00	yes	Other
7646232	https://wcids1.wixsite.com/btword	http://www.phishtank.com/phish_detail.php?phish_id=7646232	2022-08-10T09:07:24+00:00	yes	2022-08-10T09:12:58+00:00	yes	Other
7646231	http://chiyuo.ec-site.net/wp/nhcb.bns/5/nssb.php	http://www.phishtank.com/phish_detail.php?phish_id=7646231	2022-08-10T09:01:17+00:00	yes	2022-08-10T09:03:18+00:00	yes	Other
7646230	https://homesite.app-online1.repl.co/validacion.php	http://www.phishtank.com/phish_detail.php?phish_id=7646230	2022-08-10T09:00:54+00:00	yes	2022-08-10T09:03:18+00:00	yes	Other
7646229	https://onlinebanking.galiciaonline13.repl.co/	http://www.phishtank.com/phish_detail.php?phish_id=7646229	2022-08-10T08:56:58+00:00	yes	2022-08-10T09:03:18+00:00	yes	Other
	https://galiciahome	http://www.phishtank.	2022-08-10T08:54		2022-08-10T09:03:		

Figure 2: dataset 4

The dataset consists of a collection of phishing website instances. Each instance contains the URL and the relevant HTML page. The index.sql file is the root file, and it can be used to map the URLs with the relevant HTML pages. The dataset can serve as an input for the machine learning process. The only dataset available up to date.

Classification Model Setup

Feature Extraction

The underlying move toward make our portrayal model will use to recognize phishing destinations is to fabricate a stamped dataset to set up the model with.

For that we truly need to have a once-over of phishing and non-phishing destinations that we will isolate a lot of components from.

The bits of knowledge concerning the features executed should be visible as here: Features.md

The course of action of phishing and non-phishing site that were the commitment to our part extraction are independently:

verified_online.json this is a JSON with an assortment of checked phishing destinations from PhishTank.com, an unprecedented stage for doing combating phishing as well, and it uncovered its data base for engineers, you can get it here

top-1m.csv this is a csv with the best 1 million accepted destinations from Alexa, this will go about as our non-phishing locales list.Directly following these two reports in the data library, its essentially an issue of running ourcomponent extraction scripts:

Extract_Features.py and Extract_Features_Non_Phish.py to remove the phishing and non-phishing dataset independently. These are extraordinarily multi-hung scripts that follow the Thread pool plan.

Feature Groups

These features could be disconnected into 4 social occasions:

URL based features: These are the components that research the URL of the site.

Peculiar based features: These components oversee servers and require the usage of untouchables like WHOIS informational collection.

HTML and JavaScript based features: For these components, we included the Web Scraping methodology to isolate data from the HTML and JavaScript code.

Space based features: These components eliminated from the WHOIS Database .

By and by, this adventure could isolate 19 unmistakable components from each site.

Feature Details

Below, we give the names of the features used by their group:

Table 1 features details

Feature Group	Features' Names
URL based features	Having IP address
	URL length
	URL having "@" Symbol
	HTTPS token in the domain part of the url
	Shortening services
	double_slash_redirecting
	Prefic_Suffix in the domain
	Having subdomains
	sum_of_symbole_eq
	sum_of_symbole_and
	exist_of_symbol_ab
	exist_of_symbole_anch
Abnormal based features	URL of Anchor
	SFH
	Port

HTML and JavaScript based features	Redirect
	IFrame
	Redirect_html
Domain based features	Age of Domain
	Domain registration duration

Below are the details of each feature, and the rules behind its extraction:

Table 2: feature wise rules and extraction

Feature Group	Feature's Name & Details
Having IP address	If the domain part has an IP address : return 1 otherwise : return -1
URL length	If URL length < 54 : return -1 else if 54<=URL length <=75 : return 0 else : return 1
URL having "@" Symbol	If having « @ » symbol then : return 1 else : return -1
HTTPS token in the domain part of the URL	If using HTTP token in domain part of the URL: return 1 otherwise return -1
Shortening services	If using shortening services : return 1 else : return -1
double_slash_redirecting	If the existing of « // » in the path part of the URL : return 1 else : return -1
Prefix_Suffix in the domain	If the domain name part including « - » symbol : return 1 else : return -1
Domain registration duration	If Domain expires on ≤ year : return 1 else if domain expires on ≥ 1 year : return -1 else record in WHOIS not existing : return 0
Having subdomains	If number of subdomains >1 : return 1 else if number of subdomain <=1 : return -1 else if (TLD non-existing in Mozilla's TLD list) : return 0
Port	If the number of port used is of the preferred status(fig1) : return -1 else : return 1
sum_of_symbole_eq	If the number of « = » symbol < 3 : return -1 else : return 1
sum_of_symbole_and	If the number of « & » symbol < 3 : return -1 else : return 1
exist_of_symbol_ab	If the URL has « ~ » symbol : return 1 else : return -1

exist_of_symbole_anch	If URL has « # » symbol : return 1 else return -1
URL of Anchor	If % of URL anchor < 31% : return -1 else if % of 31%≤ URL anchor ≤ 67%: return 0 else : return 1
Redirect	If the website has been redirected less then 2 times : return -1 else if it has been redirected twice : return 0 else : return 1
Redirect_html	If the number of redirections in the HTML DOM =0 : return -1 else : return 1
IFrame	If using Iframe : return 1 else return 0
Age of Domain	If age ≥ 6 months : return -1 else if no WHOIS record : return 0 else : return 1
SFH	If the SFH contains empty string or "about:blank" : return 1 else if SFH doesn't exist : return 0 else : return -1

After the past step we should have 2 new records under data called extracted_Non_Phish.csv and extracted_Phish.csv that will go about as the commitment to our model arrangement:

classifier.py has the pipeline to get ready and test then, dump the model to classifier.pkl python object that will be used after to affirm the URL entered to PhishRod. It in like manner has a section to cross endorse the model and picture its different parts like part importance. So expecting any movements ought to be brought to PhishRod request model it should live there.

To run the classifier script on have the model, recently run:

the result should be the model dump at classifier/classifier.pkl so guarantee it exists preceding moving to the ensuing stage.

Tha point of collaboration will be simple, with one data district to enter the URL and a button to perceive whether or not the site is a phish. While investigating the client will find a couple of bits of knowledge around late tests, and a technique for sending analysis and contact the admin. After Sending a URL the client will have the results not long later:

Phish not perceived: a lock development will appear, and a rating space for the client to send his rating on the results

References

[1] Forbes (2019 (accessed October 11, 2020)) Artificial intelligence examples. https://www.forbes.com/sites/bernardmarr/2018/04/30/27-incredible-examples-of-ai-and-machinelearning-in-practice/#67c9eb217502

[2] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y (2020) Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2(4):230–243

[3] Barbash GI (2021) New technology and health care costs-the case of robot-assisted surgery. The New Engl J Medi 363(8):701

[4] Burgio LD, Allen-Burge R, Roth DL, BourgeoisMS, Dijkstra K, Gerstle J, Jackson E, Bankester L (2021) Come talk with me: Improving communication between nursing assistants and nursing home residents during care routines. The Gerontologist 41(4):449–460

[5] Erickson T, Danis CM, Kellogg WA, Helander ME (2008) Assistance: the work practices of human administrative assistants and their implications for it and organizations. In: Proceedings of the 2022 ACM conference on Computer supported cooperative work, ACM, pp 609–618

[6] Brady M (2021) Artificial intelligence and robotics. Artif Intell 26(1):79–121

[7] F5 (2020 (accessed April 20, 2020)) Report by f5 labs. https://www.darkreading.com/attacksbreaches/new-report-iot-now-top-internet-attack-target/d/d-id/1333147

[8] Kumaraguru P, Rhee Y, Acquisti A, Cranor LF, Hong J, Nunge E (2022) Protecting people from phishing: the design and evaluation of an embedded training email system. In: CHI 2021: proceedings of the SIGCHI conference on human factors in computing systems, ACM, New York, pp 905–914

[9] Sheng S, Magnien B, Kumaraguru P, Acquisti A, Cranor LF, Hong J, Nunge E (2022) Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish.
In: SOUPS 2022: proceedings of the 3rd symposium on usable privacy and security, ACM,
New York, pp 88–99

[10] Sheng S, Wardman B, Warner G, Cranor LF, Hong J, Zhang C (2020) An empirical analysis of phishing blacklists. In: CEAS 2009

[11] Almomani A, Gupta BB (2020) Phishing dynamic evolving neural fuzzy framework for online detection zero-day phishing E-mail. IJST 6(1):122–126

[12] Zhang Y, Hong JI, Cranor LF (2020) Cantina: a content-based approach to detecting phishing web sites. In: Proceedings on WWW, ACM, New York, pp 639–648

[13] Chen K-T, Huang C-R, Chen C-S (2021) Fighting phishing with discriminative key point features. IEEE Internet Community

[14] Ferrara, Emilio. "The History of Digital Spam". In: Commun. ACM 62.8 (July 2020), pp. 82–91. ISSN: 0001•0782. DOI: 10.1145/3299768. URL: https://doi.org/10.1145/3299768.

[15] Ludl, Christian et al. "On the Effectiveness of Techniques to Detect Phishing Sites". In: (2021). Ed. by Bernhard M. Hämmerli and Robin Sommer,
pp. 20–39.

[16] Oest, Adam et al. "Sunrise to Sunset: Analyzing the End-to-end Life Cycle and Effectiveness of Phishing Attacks at Scale". In: 29th USENIX Security Symposium (USENIX Security 20). USENIX Association, Aug.
2020, pp. 361–377. ISBN: 978•1•939133•17•5. URL: https://www.usenix.org/conference/usenixsecurity20/presentation/oest-sunrise.

[17] Siadati, Hossein et al. "Measuring the Effectiveness of Embedded Phishing Exercises". In: 10th USENIX Workshop on Cyber
Security Experimentation and Test (CSET 17). Vancouver, BC: USENIX
Association Aug. 2021. URL: https: //www.usenix.org/conference/cset17/workshop-program/presentation/siadatii.

[18] Alghamdi, H. "Can Phishing Education Enable Users To Recognize Phishing Attacks?" In: Masters dissertation, Technological University
Dublin (2021). DOI: 10.21427/D7DK8T. URL: https://arrow.tudublin.ie/scschcomdis/99/.

[19] De Kimpe, Lies et al. "You've got mail! Explaining individual differences in becoming a phishing target". In: Telematics and Informatics 35.5 (2018),
pp. 1277–1287. ISSN: 0736•5853. DOI: https://doi.org/10.1016/j.tele.2020.02.009. URL: https://www.sciencedirect.com/science/article/pii/S0736585317304677.

[20] Polit, Denise F and Beck, Cheryl Tatano. Nursing research: Principles and methods. Lippincott Williams

& Wilkins, 2004, pp. 19, 350–251.
[21] Sjouwerman, Stu. "Q4 Work From Home Phishing Emails on the rise". In:
(2021). URL: blog.knowbe4.com/infographic- q4- 2020- work- fromhome-phishing-emails-on-the-rise.
[22] Dataset 1:https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning/download?datasetVersionNumber=1
[23] Dataset 2: https://www.kaggle.com/datasets/shashwatwork/web-page-phishing-detection-dataset/download?datasetVersionNumber=2
[24] Dataset 3: https://www.kaggle.com/datasets/eswarchandt/phishing-website-detector/download?datasetVersionNumber=3