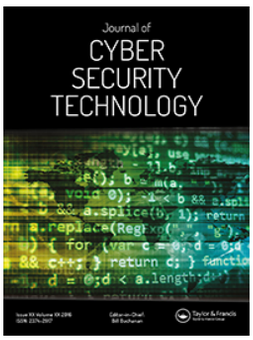


Journal of Cyber Security Technology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tsec20

Phishing website detection based on effective machine learning approach

Gururaj Harinahalli Lokesh & Goutham BoreGowda

To cite this article: Gururaj Harinahalli Lokesh & Goutham BoreGowda (2020): Phishing website detection based on effective machine learning approach, Journal of Cyber Security Technology, DOI: 10.1080/23742917.2020.1813396

To link to this article: https://doi.org/10.1080/23742917.2020.1813396

Published online: 31 Aug 2020.

JOURNAL OF CYBER SECURITY TECHNOLOGY
https://doi.org/10.1080/23742917.2020.1813396

Taylor & Francis
Taylor & Francis Group

Phishing website detection based on effective machine learning approach

Gururaj Harinahalli Lokesh and Goutham BoreGowda

Wireless Inter Networking Research Group (Wing), Vidyavardhaka College of Engineering, Mysuru, India

ABSTRACT

Phishing a form of cyber-attack, which has an adverse effect on people where the user is directed to fake websites and duped to reveal their sensitive and personal information which includes passwords of accounts, bank details, atm pin-card details etc. Hence protecting sensitive information from malwares or web phishing is difficult. Machine learning is a study of data analysis and scientific study of algorithms, which has shown results in recent times in opposing phishing pages when distinguished with visualization, legal solutions, including awareness workshops and classic anti-phishing approaches. This paper examines the applicability of ML techniques in identifying phishing attacks and report their positives and negatives. In specific, there are many ML algorithms that have been explored to declare the appropriate choice that serve as anti-phishing tools. We have designed a Phishing Classification system which extracts features that are meant to defeat common phishing detection approaches. We also make use of numeric representation along with the comparative study of classical machine learning techniques like Random Forest, K nearest neighbours, Decision Tree, Linear SVC classifier, One class SVM classifier and wrapper-based features selection which contains the metadata of URLs and use the information to determine if a website is legitimate or not.

ARTICLE HISTORY
Received 15 April 2020
Accepted 14 August 2020

KEYWORDS-
Machine learning; phishing; legitimate; random forest classification

1. Introduction

Machine learning is a multidisciplinary approach initially used in supervised learning to form analytical models. It plays a major aspect in a broad scope of serious applications such as image recognition, data mining, skilled systems and image recognition. This approach appears suitable to solve phishing page detection, because this problem can be converted into a task of classification. ML techniques can be used to develop models to detect phishing activities based on categorizing old web pages and then these models can be integrated into the browser. Consider an example of a user browsing a web page, ML models will find

CONTACT Gururaj Harinahalli Lokesh gururaj1711@vvce.ac.in Wireless Inter Networking Research Group (Wing), Vidyavardhaka College of Engineering, Mysuru, Karnataka, India

the legitimate website instantly and then forward the output to the user at the other end. The vital factor for the success is the website's features in the input dataset and the availability of adequate websites for the creation of trustworthy analytical models, in developing ML models for automated anti-phishing identification [1,2].

We already learnt that, Phishing is a cyber-attack in which a person is made to visit illegal websites and fooled to reveal their hypersensitive data like name of user, bank details, card details, passwords etc. As primary security really matters on the web, phishing has drawn consideration of many experts and researchers. When there are two similar web pages, and information accompanied to the first page on apprehensive is entered by the user, an alert message should be raised on the second page second. When two web pages are not same, it is absurd that legitimate site is spoofed by second page, and thus the information can therefore be passed on without an alert that the page obtained is a legitimate page, based on keywords, by search done using a search engine or choosing between a set of predefined registered pages[2, 3, 4].

There are tools, capital of literature and methods for serving web users to recognise and refrain from phishing web pages. Some of the present phishing identification techniques are skilled in detecting phishing webpages with an extreme accuracy (>99%) while attaining extremely low accuracy of false classifying legitimate webpages (<0.1%). Although, a large number of these techniques, which make use of machine learning mainly depends on lots of inert characteristics, chiefly using the bag-of-words approach. As phishing identification methods struggle with gaining and upholding labelled data of training dataset. In accordance with deplorability perception, solutions which accordingly need minimum data for training are thereby very appealing [1,5,6]. Because of unavoidable phishing web pages mainly aiming at banks, online trading, governments and users of the web, it is necessary to avoid phishing attacks of web pages at the initial phase. Although, identification of a phishing web page is a laborious task, by virtue of the number of advanced approaches used by attackers to step out users of the web. The triumph of phishing web page identification techniques chiefly rely on identifying phishing web pages precisely and within an adequate period of time. As substitute solutions to the predictable phishing web page identification methods, a few inventive phishing identification methods are established and proposed in order to efficiently foresee phishing web pages. Over the last few years, the exceptional phishing web pages detection methods based on controlled machine learning techniques have been more often, which are more adaptive and clever to the atmosphere of the web associated with the predictable phishing web page identification methods [6].

The motivation in taking up the work is due to increasing phishing attacks from day to day and during the covid-19 pandemic it has doubled in numbers. According to the McAfee Covid-19 Threat Report, cyber criminals have been exploiting the pandemic through coronavirus-related malicious apps, phishing

campaigns and malware, focusing on topics such as testing, treatments, cures and remote work. KnowBe4 reveals 56% of simulated phishing tests were related to coronavirus. Social media messages are another area of concern when it comes to phishing. Within the same report, KnowBe4's top-clicked social media email subjects reveal password resets, tagging of photos and new messages. Another example is the online classes taken on various video call platforms where there is a high chance of someone posting an unknown link which might lead to phishing.

In this paper we make use of Random forest algorithm which is a collective learning technique for regression, classification and other tasks that works by creating an assembly of decision trees in a training set and ensuing in a class that is a mean prediction of the individual trees or the mode of the classes. The universal technique of random decision forests was first proposed by Tin Kam Ho in 1995. He emphasizes that forests of trees piercing with sloping hyperplanes as they can gain accuracy as they grow without being affected from overtraining, as long as the randomly limited forests are to be sensitive to only selected dimensions. The observation of a more complicated classifier obtained a more precision of monotonically sharp distinction to a collective belief that the complication of a classifier can solely raise to a point of accuracy before offended by over fitting.

This article follows the following structure: Section 2 describes the Background and Existing Systems, Section 3 is the description about the dataset that we used, Section 4 and 5 are the case study analysis and the technique that we have implemented in our work, and it also includes mathematical models. Section 6 is the conclusion and then the references.

2. Background and existing system

In [7], they have developed a system that measures the conduct of the social architects, and a complete model for depicting mindfulness, estimation and resistance of social building-based assaults. They have proposed a hybrid multi-layered model utilizing normal language handling strategies for guarding the social designing-based assaults. The show empowers the fast recognition of a potential assailant attempting to control the unfortunate casualty for uncovering secrets. In this model they make use of a model named Security Training and Processing Evaluation (STPE) and this model contains a cycle with five stages. This model helps to protect the sensitive information from social engineering attacks.

In another method, they make use of a phishing location and anticipation method by joining URL-based and web page by similitude-based discovery. URL-based recognition includes selection of genuine URL (to which the site is actually coordinated) and the visual URL (which is identified by the client). This paper detects the phishing sites in two phases. The first phase is URL and Domain Identity Verification, in this phase we make use of LinkGuard algorithm to inspect

the two URLs and then based on the result the procedure will proceed to the next stage. The second phase is image-based page matching. In this phase a snapshot of the original webpage and the suspicious web page will be taken, this is done either by the code developed or by utilizing a browser plug-in for webpage snapshot. Then they compare the snapshots, first they modify the image so that we have only less comparisons. They applied various transform methods like DFT (Discrete Fourier Transform), DCT (Discrete Cosine Transform) and other techniques like cross-correlation. If the detection of phishing sites are not detected by URL-based detection, then we make use of visual similarity-based detection. One of the novel techniques to check the site is legitimate or not [8].

In other research work, the proposed system used secure QR code as an Anti-Phishing mechanism to stop web phishing. The system depends on the image captcha acceptance plan utilizing visual cryptography. It expects key and secret data from the phishing sites [9].

Waleed Ali proposed a procedure for detecting phishing websites by making use of supervised machine learning techniques such as radial basis function network (RBFN), naïve Bayes classifier (NB), back-propagation neural network (BPNN), decision tree, k-Nearest neighbour (kNN), random forest (RF) and support vector machine (SVM) a technique of detecting phishing website with wrapper features selection based on machine learning classifiers. In the research conclusions, the Neural Networks model was used in the process of classification, but it was prone to under fitting because it was poorly structured [10]. However, it would over fit the training data set if structured to each single item in the dataset [11,12].

In this experimentation which is based on a number of features of the dataset which reveals that the self-structuring NN model was able to generate highly predictive anti-phishing models compared to other traditional C4.5 and probabilistic classification approaches [1].

The features which were considered include images, text pieces and styles, signature extraction, URL keywords and the overall appearance of the page as rendered by the browser were identified and considered for the experiment [3].

3. Dataset description

Main challenge we faced was to find legitimate datasets for the model. Many researchers face the same problem while working in this field. Thus, it was very burdensome to find a dataset that fulfils all required features. Datasets used for the research purpose are collected mainly from MillerSmiles archive and Phish Tank archive which are extracted using data mining algorithms. In the dataset, features extracted were achieved manually, but the human interaction with the system plays a vital role which might affect the exposure to phishing attacks. The dataset used for this work is taken from the link provided here: https://archive.ics.uci.edu/ml/machine-learning-databases/00327/.

3.1. Dataset representation

Dataset contains new features which also contain experimentally one where new rules are assigned to some well-known parameters. There are 30 parameters in the dataset and has been listed below:

'having-IP-Address','Prefix-Suffix', 'having-Sub-Domain', 'SSLfinal-State', 'Domain-registration-length', 'Favicon', 'port', 'URL-Length', 'Page-Rank', 'Google-Index', 'Shortening-Service', 'having-At-Symbol', 'double-slash-redirecting', 'DNSRecord', 'web-traffic', 'Links-pointing_to_page', 'Statistical-report', 'Abnormal-URL', 'Redirect',URL-of-Anchor', 'Links-in-tags', 'on-mouseover', 'RightClick', 'popUpWidnow', 'Iframe', 'age-of-domain', 'HTTPS-token', 'Request-URL', 'SFH', 'Submitting-to-email'.

This model determines whether it a phishing site or not based on the following stages:

3.1.1. Address bar features

- Presence of IP address in the URL: If the link has an IP address as a part of it, then it is treated as a phishing site, else it is treated as legitimate site.
- Length of URL: Average length of legitimate site is less than or equal to 54 [13]. If the length of the link is less than 54 then it is a vulnerablesite or if the length is greater than 54 but less than 75 then it is treated as a suspicious site else it is treated as a phishing site.
- Using URL shortening services: If the link uses such service, then it is treated as a phishing site or else it is treated as legitimate site.
- Presence of '//' in the link: This means that it will redirect the user to another website. But every URL has a '//' after the specified protocol (Ex; HTTP, HTTPS). So, if '//' appears after the seventh position, then it is treated as a phishing site, else it is treated as a legitimate site.
- Presence of sub domains and multiple sub domains in the link: When URL has no sub domains, then it is treated as a legitimate site. But most websites have sub domains, if the number of periods encountered is greater than one (excluding www.), then the URL is regarded as suspicious. However, if the periods are greater than two, then it is a phishing site.
- Existence of HTTPS protocol: If the website uses HTTPS protocol, its certificate is issued by trusted party, and age of certificate is valid then, it is a legitimate site. If a certificate is provided by a party which is not trusted, then it is suspicious or else it is a phishing site.
- Favicon image associated with the website: If the graphic image is loaded from a domain which is not from that of the given link, then that web page is considered as a phishing site.
- Usage of nonstandard port numbers: If the port no. for services running in the server is different for the website than the standard port number specified, then it is a phishing site.

3.1.2. Abnormal-based features

- Existence of request URL and anchor tag: Request URL checks whether the external objects are queried from another domain and average percentage allowed is 22% [13]. If the percentage is less than 22%, then it is a legitimate site or else it is greater than 22% but less than 61%, it is categorized as suspicious else it is treated as a phishing site.
- Using <Meta>, <Script> and <Link> tags: Average percentage of anchor tag present is a site is 31% [13]. If the percentage is less than 31%, then it is a legitimate site or else it is greater than 31% but less than 67%, it is treated as suspicious else it is categorised as a phishing site.
- Server Form Handler(SFH) webpage: SFHs redirecting to different domain names of the that of given link which might contain about:blank or an empty string are doubtful because action takes place after the information is submitted. If the SFHs is about:blank or IsEmpty then is a phishing site or else if it requests another domain, it is suspicious, else it is a legitimate site.
- Website that submits information to Email: Forms on the website always submits information to a server for processing, but the attacker redirects the information to his database. If mailto() or mail() function is used to submit user information on a site, then it is a phishing site.
- Domain registered in WHOIS database: If the hostname present in the URL is not registered under WHOIS, then is treated as a phishing site.

3.1.3. HTML and javascript-based features

- Website Redirected Count: On average, a legitimate site redirects 1 time and phishing site redirects at least 4 times [13]. If the site redirects more than 2 but less than 4 times, then it is a suspicious site. If it redirects more than 4 times, it is a phishing site.
- Customization of status bar: Attackers may fake the URL displayed on the status bar. Hence, onMouseOver function is used to detect the change and flag it as a phishing site.
- Disabling Mouse events: Attackers disable the right click by using JavaScript to prevent the users from opening the source code to verify. So, if eventbutton = = 2 is present which disables right click, then it is a phishing site.
- Frequent Popup windows: No legitimate site uses a pop-up window to ask users to submit information. If the pop-up window prompts for a form asking for information, then it is categorised as a phishing site.
- Iframe redirection existence: IFrame, a HTML tag used to display another page into the current one. Attackers take advantage of it to make current pages invisible by displaying phishing pages without frame borders. Those links are classified as phishing sites.

3.1.4. Domain-based features

- Lifetime of Domain: Expired validity of the domain present in the link, then it is considered as a phishing site.
- DNS record: If the DNS record is unavailable in WHOIS database, then it is categorised as a phishing site.
- No. of visitors to the webpage: Alexa database holds information about websites and genuine websites are ranked among the top 1,00,000 [13]. Further, if there is no traffic or the domain is not found in the Alexa database, then it is treated as a phishing page.
- Rank of the webpage: PageRank is an algorithm used by Google Search to rank websites and there are no PageRank for 95% of phishing webpages [13]. If the PageRank value is less than 0.2, then it is a phishing site.
- Google Index of the webpage: Google index is not provided for any website in short span. So, if the website is not indexed by Google, then it is classified as a phishing page.
- No. of links pointing to the page: Genuine websites have at least two external links pointing to them and 98% of phishing pages have no links cited to them [1]. If there are no links pointing to them, they are treated as phishing links or else they are categorised as suspicious links, if no. is greater than 0 but less than 2.
- Report on the website: PhishTank and StopBadware are open source popular websites which house data and information about phishing websites on the internet. If the links are flagged as phishing sites on their website, then they are phishing links.

4. Case study analysis

The system specifications used for this project is Intel core i5 with 8GB RAM and 5GB free hard disk space. It was performed on GNU/Linux (can also be performed on WIndows/Mac OS). Project is written in Python using its libraries in Jupyter Notebook. Alternatively, we can use Google Colab service to implement the project.

4.1. Random forest classifier

A supervised machine learning algorithm random forest can perform both classification and regression tasks. Classification helps to classify our data for categorical variables. Regression helps to predict outcome of data for example to predict the salary of a person based on their experience.

Random Forest is an ensemble-based technique. Ensemble algorithms combine two or more algorithms of the same or diverse kind to classify objects. When a random forest classifier is applied first it will pick a random K data point from the training dataset and then build a decision tree associated with each of these data points. Then we can choose the 'N' number of trees we need to

perform the first step repeatedly. Atlast for a new datapoint, make each and every 'N' number of trees to anticipate the category to which it belongs, and allocate the new datapoint to the category that has the maximum vote.

Basic parameters that can be taken in a random forest classifier is the total count of trees to be generated ie., n-estimators by default this parameter will take value as 10. Then the parameter max-depth specifies the maximum depth of the tree. It is by default set to none if this parameter is not specified. If none then nodes are extended till the leaves are absolute. Next prominent parameter is max-leaf-nodes. This parameter is used to grow the trees with max-leaf-nodes in best-first fashion. By default this also takes as none which means unlimited number of leaf nodes.

4.2. Decision tree classifier

Decision Trees are used for regression and classification purposes and its a non-parametric supervised learning method. Decision Tree classifier will produce a model that prophecies the estimation of target variable by learning rules of decision which are inferred from the data features. Decision tree algorithms are associated with a set of if-then-else decision rules. If the tree is deeper, then the decision rules are more complex and the model is better fitter. Decision tree classifiers build tree-like structure models.

The algorithm splits the dataset into smaller subsets and the related decision tree will be enhanced simultaneously. The obtained result will finally be a tree which consists of leaf nodes and decision nodes. A classification or decision is represented by the leaf nodes. A decision node is the node which will have branches that are two or more in number. The highest decision node in a tree which corresponds to the finest predictor is called the root node. Both numerical and categorical data are handled by the decision trees.

4.3. K nearest neighbours

K-Nearest Neighbour (kNN) is a non-parametric supervised machine method. The working of the kNN algorithm is as mentioned. Whenever a new datapoint is to be added to the model to classify to which category the new datapoint it belongs to first it will choose the number of neighbours (k) and then it will take the K nearest neighbours of the new datapoint based on the Euclidean distance(or any other method specified in the parameter). Among these K neighbours, it will count the total statistics of data points in each category. At last it will allocate the new datapoint to the category where the count has more in the counted categories. To get more accuracy we can vary the value of K.

4.4. Linear SVC classifier

The Linear Support Vector Classifier(SVC) is used to fit to the data that has been provided. A best fit hyper plane will be returned after applying SVC classifier to a dataset. And this hyperplane will divides, or classifies, dataset in best fit fashion. Once the hyperplane has been obtained we will upload some capabilities to the classifier to look at what the anticipated class is.

4.5. One class SVM classifier

The support vector machine (SVM) is amongst the most notable and powerful techniques in supervised machine learning. We can also perform classification tasks in data mining using SVM. The working of SVM is as explained. To create a boundary between the different classes of the dataset it will generate a hyperplane. To choose a hyperplane there are certain criteria's to be satisfied. The hyperplane separates the different classes and it should maximizes the margin (means it is a distance from point that is nearest to the hyperplane) with the different type of classes. A boundary will be obtained between the various classes upon creating the hyperplane, a boundary. Finally, we are able to characterize any data to a class by identifying the class to which the data point belongs to.

The first task is to divide the dataset into training dataset and testing dataset, in the proportion of 80:20. The purpose of training a dataset is to train different models, and the trained models are fed with a test dataset to check the result.

First of all, Linear SVC classifier was applied on the dataset by the default parameter values and the kernel type with linear. The accuracy obtained from this classifier is 92.69%. After this, Decision tree classifier was applied with default parameter values and 96.05% accuracy was obtained. Then the K-Nearest Neighbour (KNN) algorithm was applied with the parameter value 5 for n_neighbours and the algorithm applied for KNN is ball_tree. With these parameter values we obtained the accuracy as 93.53%. For One Class SVM classifier we obtained the accuracy rate of 48.56%. Finally when the Random forest classifier was applied on this dataset with the values of n_estimators as 500, max_depth as 15 and max_leaf_nodes = 10,000, we obtained the highest accuracy rate of 96.87%. Thus, the efficacy of Random Forest is better than the rest of the algorithms. The outcomes for the dataset are outlined in Table 1.

5. Random forest algorithm

The algorithm of Random forest is split into two stages: Creation and Prediction.

Creation Algorithm:

(1) In random choose 'f' features from the total 'f_t' features in which $f << f_t$
(2) Among all the 'f' features, node 'n' is calculated using the best split point.

Table 1. Accuracy of different algorithms for the dataset taken.

Algorithm	Accuracy
One Class SVM	48.56%
Linear SVC classifier	92.69%
K-Nearest Neighbour	93.53%
Decision tree classifier	96.05%
Random Forest	96.87%

(3) Daughter nodes are created by splitting the node using the best split.
(4) Repeat steps 1 to 3 till the 'l' number of nodes is reached.
(5) Forest is built by repeating steps 1 to 4 for 't' no. of times to form 't' number of trees.

Prediction Algorithm:

(1) Result is concluded using the rules of the decision tree by taking the analysed features randomly and stores the result.
(2) Votes are calculated for each and every specific target
(3) Contemplate the highest voted target as the concluded indicator from the algorithm.

Choose **'f'** features from **'f_t'** using the best split approach to discover the root node of the tree. Later, the same technique is used to calculate daughter nodes of the tree. First three stages are continued till the tree with a leaf node and root node having a target is formed. At last, one to four stages are repeated to form **'t'** randomly formed trees thus forming the random forest. Test features are passed to the trained algorithm for every randomly chosen tree and votes will be computed for unique targets out of total trees. The one with the highest vote will be considered as the resultant value and this process is called voting for the majority.

5.1. Mathematical model

As a gist, this algorithm is a collection of correlated decision trees. It creates many decision trees, which helps in the classification based on bagging technique.

Consider amatrix S, with training examples, fed to the algorithm to create a classification model as depicted in Figure 1.

In this case, elements are features of the dataset where f_{A1}, f_{B1} and f_{C1} are feature A, B and C of the 1st sample respectively and so on whereas C_1 and C_N are the training class of the respective sample to classify the sample set.

Consider the random sample set from the forest as shown in Figure 2,

Each subset is a collection of different features and from these subsets, each decision tree is created from the respective matrix S_1, S_2 so on up to S_M. After the

$$S = \begin{bmatrix} f_{A1} & f_{B1} & f_{C1} & C_1 \\ \vdots & & \vdots & \\ f_{AN} & f_{BN} & f_{CN} & C_N \end{bmatrix}$$

Figure 1. Matrix S.

$$S_1 = \begin{bmatrix} f_{A12} & f_{B12} & f_{C12} & C_{12} \\ f_{A15} & f_{B15} & f_{C15} & C_{15} \\ \vdots & & \vdots & \\ f_{A35} & f_{B35} & f_{C35} & C_{35} \end{bmatrix} S_2 = \begin{bmatrix} f_{A2} & f_{B2} & f_{C2} & C_2 \\ f_{A6} & f_{B6} & f_{C6} & C_6 \\ \vdots & & \vdots & \\ f_{A20} & f_{B20} & f_{C20} & C_{20} \end{bmatrix}$$

$$S_M = \begin{bmatrix} f_{A1} & f_{B1} & f_{C1} & C_1 \\ f_{A9} & f_{B9} & f_{C9} & C_9 \\ \vdots & & \vdots & \\ f_{A12} & f_{B12} & f_{C12} & C_{12} \end{bmatrix}$$

Figure 2. Matrix S_1, S_2, …., S_M.

trees are created, every tree is asked to predict its outcome based on their subset features. Now, votes are accounted from every tree and the outcome with the highest number of votes will be the result and that outcome is called the predicted class of the classifier.

5.2. Feature importance

Feature importance explains which feature played a vital role in predicting the feature class. Although the training algorithm treats every feature equally, further dataset can be processed before entering into the algorithm to increase the accuracy. Below Figure 3 shows relative significance of every independent variable present in the dataset to the classifier.

5.3. Sample tree from forest

As random forest is one among the collection of several decision trees, plotting them gives a gist of what model is trying to predict and the target values inferred from it. Visualizing a few trees will provide a good intuition about the model.

In the figure below An Effective Machine Learning Based Detection, feature name, split value, splitting criteria used(default 'gini'), no of samples, no of samples of each class is visualized.

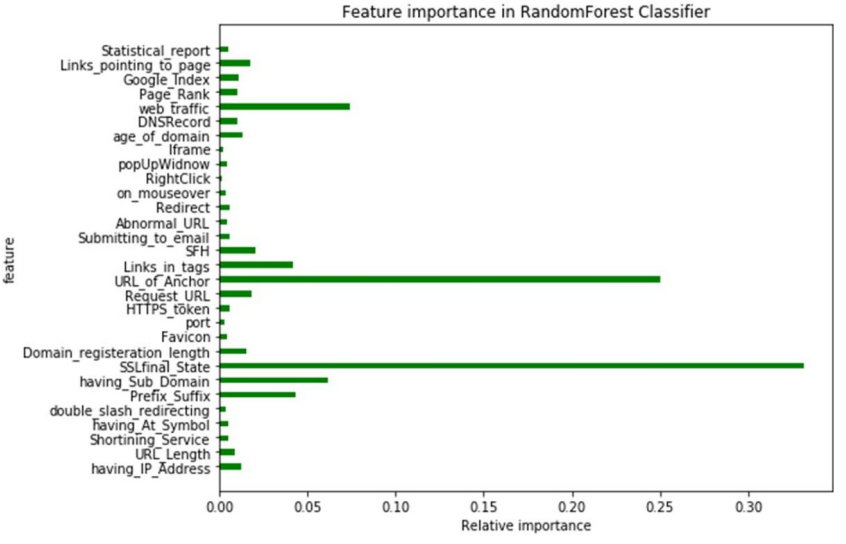
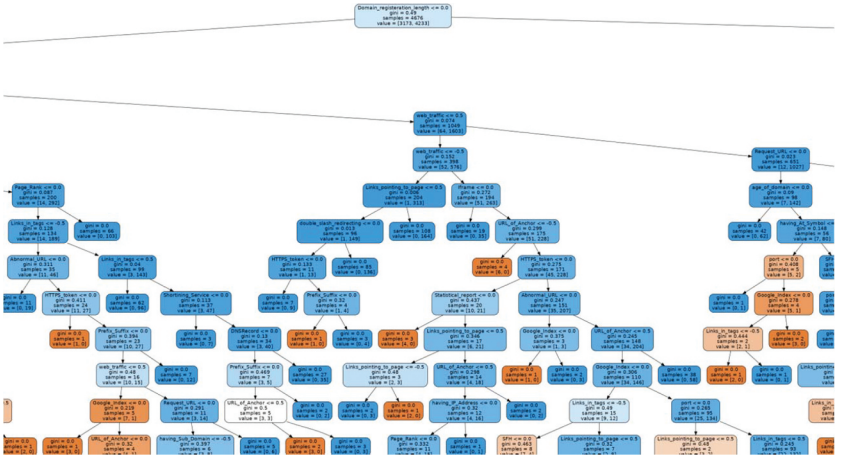


Figure 3. Feature importance in random forest classifier.



Complete sample tree can be viewed in the following link: https://drive.google.com/file/d/1iwZgGwi0Ewbu_Pel3tWF7ZL6Lto3XSwO/view?usp=sharing

6. Conclusion

Phishing is a critical menace to users data nowadays. Detection of phishing websites is a tedious job, as the result phishers are rapidly increasing. To overcome the issue, researchers and experts worked on many approaches and techniques, but it resulted in low rates of detection. For our work, we used many techniques such as Decision tree Classifier, K nearest neighbours, Linear SVC classifier, Random Forest classifier, One class SVM classifier. Out of which we

observed that Random Forest got the highest accuracy of about 96.87% when compared with other methods as listed in Table1. Whereas one class SVM becomes the one with least accuracy of about 48.56% .We split it into stages: Creation and Prediction, these algorithms used to build the forest and predict the results as explained previously. We predominantly observed that Random Forest performed better than other methods or algorithms as mentioned above. Overfitting of data is avoided, which is one of the important feature. Hence Random Forest classifier is best suited for us to detect more accurately whether the website is phishing or not.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Dr. Gururaj Harinahalli Lokesh is currently working as Associate Professor, Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, India. He holds a Ph.D. Degree in Computer Science and Engineering from Visweswaraya Technological University, Belagavi, India in 2019. He is a professional member of ACM and working as ACM Distinguish Speaker from 2018. He is the founder of Wireless Internetworking Group(WiNG). He is a Senior member of IEEE and lifetime member of ISTE and CSI. **Dr. H L** received young scientist award from SERB, DST, Government of India in Decemeber 2016. He has 9 years of teaching experience at both UG and PG level. His research interests include Block Chain Technology, Cyber Security, Wireless Senor Network, Ad-hoc networks, IOT, Data Mining, Cloud Computing and Machine Learning. He is an Editorial Board member of the International Journal of Block chains and Cryptocurrencies (Inderscience Publishers) and Special Editor of EAI publishers. He has published more than 75 research papers including 2 SCI publications in various international journals such in IEEE Access, Springer Book Chapter, WoS, Scopus, and UGC referred journals. He has presented 30 papers at various international Conferences. He has authored 1 Book on Network Simulators. He worked as reviewer for various journals and conferences. He also received Best paper awards at various National and International Conferences. He was honored as Chief Guest, Resource Person, Session chair, Keynote Speaker, TPC member, Advisory committee member at National and International Seminars, Workshops and Conferences.

Goutham. B completed his B.E in Electrical and Electronics Engineering from Visvesvaraya Technological University, Belgaum India in 2013 and M.Tech degrees in Computer Application in Industrial Drives from Shri Siddartha Academy of Higher Education ,India in 2015. Currently he is working as an Assistant Professor in the Department of Electrical and Electronics Engineering at Vidyavardhaka College of Engineering Mysuru, India. His areas of interest include Smart Grids, Cyber Security, MicroGrids, Renewable Energy sources, Electrical Machines.

ORCID

Gururaj Harinahalli Lokesh http://orcid.org/0000-0003-2514-4812

References

[1] Abdelhamid N, Thabtah F, Abdel-jaber H Phishing detection: a recent intelligent machine learning comparison based on models content and features. In Beijing, China: IEEE; 2017.

[2] Harikrishnan NB, Vinayakumar and Soman KP on "A machine learning approach towards Phishing email detection; 2018.

[3] Damodaram R, Valarmathi ML Phishing detection based on web page similarity. In IJCST; 2011.

[4] Jagadeesan,Anchit S, Chaturvedi and Kumar S. URL phishing analysis using random forest. Int J Pure Appl Math. 2018. 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS), Page No. 1063-6927, Nara, Japan.

[5] Marchal S, Saari K, Singh N, et al. Know your phish: novel techniques for detecting phishing sites and their targets. arXiv. 2016.

[6] Ali W. Phishing website detection based on supervised machine learning with wrapper features selection. Int J Adv Comput Sci Appl. 2017 September;8(9). DOI:10.14569/IJACSA.2017.080910.

[7] Thakur K, Shan J, Pathan A-SK .Innovations of phishing defense: the mechanism, measurement and defense strategies. In International Journal of Communication Networks and Information Security (IJCNIS); 2018 April 1.

[8] Shekokar NM, Shah C, Mahajan M, et al. An ideal approach for detection and prevention of phishing attacks.

[9] Shaikh R, Mala S, Salman A, et al. A mobile based anti-phishing scheme using QR code. In: International Journal of Innovative Research in Computer and Communication Engineering; 2016 October 10.

[10] Duffner S, Garcia C, An online backpropagation algorithm with validation error-based adaptive learning rate. In: Artificial Neural Networks – ICANN 2007; Porto, Portugal; 2007.

[11]. Mohammad RM, Thabtah F, McCluskey L. Predicting phishing websites based on self-structuring neural network. Neural Comput Appl; 2014.

[12] Thabtah F, Mohammad RM, McCluskey L. A dynamic self-structuring neural network model to combat phishing. 2016 International Joint Conference on Neural Networks (IJCNN), Canada; 2016.

[13] Mohammad R, McCluskey TL, Thabtah F An assessment of features related to phishing websites using an automated technique. In: International Conference For Internet Technology And Secured Transactions. London, UK: ICITST; 2012.