HIGH ACCURACY PHISHING DETECTION

Kushal G Krishna

Kushal G Krishna, Bangalore, Karnataka 560072
Student, Bangalore Institute of Technology, Dept of Information Science and Engineering, VV Puram, Bangalore, Karnataka 560004

---***---

Abstract - *One kind of internet security issue, phishing websites, prey on users' natural distrust of websites and may even trick them into divulging personal information. Essentially, it's the practise of luring victims online so that criminals may steal personal information like login credentials.*

Because of the Internet's rapid growth, phishing has become a major problem. Phishing attackers are continually developing new techniques to fool users into disclosing personal information. Because of this, a real-time anti-phishing system and an intelligent phishing detection solution are needed. Here, we develop a reliable detection system that can quickly adjust to novel circumstances and phishing domains. We use an online, feature-rich machine learning engine to differentiate between phishing and legitimate websites. The proposed approach is a client-side fix that needs no third-party support, since it just needs access to the HTML of URLs and web pages in order to gather a wide variety of discriminatory data.

As part of this initiative, we provide a smart algorithm to identify potential phishing websites. Machine learning, and more specifically supervised learning, is the basis of the system. Since Logistic Regression is so effective at categorization, we've decided to use it. The goal of our research is to find the optimal feature combination from which to train a classifier with the highest possible performance, with a particular emphasis on phishing websites.

Key Words: Cloud computing, Phishing website, machine learning, regression, reliable detection system, adaptive lerning.

1.INTRODUCTION

One kind of internet security risk, phishing websites, targets users' emotional vulnerabilities rather than their technical ones. The term "social engineering" is often used to describe the process of tricking internet users into giving up personal information like their login credentials.

Phishing is now one of the most common online hazards because of the tremendous expansion in volume that the World Wide Web has seen over the course of recent decades. Phishing attackers constantly deploy innovative strategies, including zero-day exploits and advanced methods, to trick online consumers. As a result, it is essential that the anti- phishing system employs both an intelligent and fast phishing detection solution in real time. In this piece, we build a reliable detection system that can dynamically modify its settings to account for the ever-shifting nature of the internet and phishing sites. We employ a web-based implementation of a robust machine learning engine that can identify phishing sites from legitimate ones. The proposed technique does not need any form of server-side intervention since it relies only on the client to extract various differentiating factors from URLs and the source code of websites.

We provide a thorough method for identifying phishing domains in this work. The framework relies on a machine learning technique, more especially supervised learning. Due to its superior performance in classification issues, Logistic Regression was selected. Our primary goal is to improve the performance of our classifier by analysing phishing website characteristics and determining which attribute combinations provide the best training results.

1.1 Problem statement

The Internet has developed into a medium for a wide range of criminal activities, from spam advertising to monetary fraud and virus distribution. The rapid growth of the World Wide Web and other Internet-based technologies has caused a change in consumer preferences away from in-store purchases and toward those conducted online or over the phone.

The vast majority of hackers currently concentrate their efforts on phishing as one of many ways for finding prospective victims online. Phishing is a kind of online fraud or identity theft in which a fake yet superficially similar to the victim's actual website is used to trick them into entering sensitive information.

Passwords, account information, credit card numbers, usernames, and passwords are just some of the sensitive data that phishers target in their attempts to steal from their victims. One kind of social engineering is "phishing," an acronym for "personally identifiable information theft through electronic mail." Messages sent to users frequently seem to come from trusted sources like social networks, auction sites, banks, online payment processors, or IT administrators, but they are really attempts to steal personal information. Phishing emails are another possible medium for such messages.

This kind of cyberattack often begins with a false email or other means of contact (such as a phone call or text message) that is meant to fool the recipient into giving over personal information. The message is masked to make it seem to have come from a trustworthy source, and the malicious websites they run are made to look like the legitimate one.

1.2 Research objectives

Develop a reliable solution that, given certain fundamental information on a website, can determine whether or not it has the potential to become a phishing website.

Train the model on a sufficient number of data sets in order to keep the accuracy level at or above 90%. Optimize the model in order to raise the accuracy level even further.

Make data visualisation options available to consumers so that they may have a deeper and more meaningful understanding of the patient's health

2. Literature Survey

A literature review, also known as a narrative review, is a sub-genre of review article and a sort of academic document that summarises the most recent discoveries in a field of study, as well as the theoretical and methodological advancements that have been made in relation to that field. Literature reviews are considered secondary sources since they do not report on new or previously unpublished material. Research in almost every academic discipline begins with an analysis of the relevant prior literature. Evaluative, exploratory, and instrumental assessments of previous work are the three primary categories of literature reviews.

In this study, we investigate the possibility of reaching consensus over the most crucial parameters for distinguishing phishing efforts. We use a technique based on the Fuzzy Rough Set (FRS) theory to identify the most salient features of three reference data sets. In order to identify phishing attempts, the relevant attributes are given to three different classifiers. The classifiers are trained on a separate out-of-sample data set consisting of 14,000 website samples to assess the FRS feature selection for developing a generic phishing detection. In the Random Forest classification approach, FRS feature selection may be used to get an F-measure of 95%. In addition, the FRS has uncovered 9 shared characteristics among the three data sets.

Using a technique developed from the Fuzzy Rough Set (FRS) theory, we extract the most salient features from three benchmark datasets. In order to recognise phishing attempts, the specified attributes are input into three different classifiers. Classifiers are trained using an independent out-of-sample data set of 14,000 website examples to assess the FRS feature selection for developing a generic phishing detection. It is possible to get an F-measure of 95% using FRS feature selection when using the Random Forest classification method. Additionally, FRS has identified nine shared features after analysing all three data sets. F-measure values close to 93% are achieved when using this universal feature set, which is on par with the FRS's effectiveness. This discovery shows that we may achieve speedier phishing detection that is also immune to zero-day assaults even without consulting any external sources, since the universal feature set does not contain any features obtained from third-party services. This is due to the fact that the universal feature set does not include any features that are built upon the infrastructure of other services. [1]

With the current state of phishing detection systems, dangers like zero-day phishing website assaults cannot be prevented. Due to these obstacles, a three-tiered assault detection approach called Web Crawler based Phishing Attack was developed.

A WC-PAD detector is a potential solution to this problem. Input parameters used to distinguish phishing and non-phishing websites include online traffics, web content, and the Uniform Resource Locator (URL). The suggested WC-PAD is experimentally examined using datasets gathered from actual phishing cases. The testing shows that the proposed WC-PAD can detect both known and novel phishing attacks with a 98.9% success rate.

Websites are categorised as phishing or non-phishing based on input criteria such as the volume of online traffic, the nature of the website's content, and the Uniform Resource Locator (URL). Using datasets amassed from actual phishing assaults, we undertake experimental investigation of the proposed WC-PAD. [2]

Anti-Phishing Working Group (APWG) statistics from December 2018 indicate a rise in phishing attacks targeting banking and payment systems. In order to escape detection, phishing URLs virtually always utilise HTTPS and often use redirects. This article presents a systematic literature evaluation of existing methods for recognising phishing websites. An evaluation of existing anti-phishing methods was deemed necessary, and their limitations were identified. Our primary contribution is that we researched previously used URL-based features and improved their definitions to better reflect the current state of affairs. Our study also adds to the literature by dissecting and analysing a systematic approach to building an anti-phishing model.

It was determined to conduct a comparative analysis of the anti-phishing technologies that are now in use, and the limits of these tools were recognised. Our most important contribution is that we studied the URL-based characteristics that were utilised in the past in order to enhance their definitions so that they better fit the present reality. Additionally, a method of developing an ant phishing model using a step-by-step process is explained in order to create an effective framework, which adds to the value of our contribution. The findings of this study, together with the researchers' suggestions for improving various systems, are presented here.[3]

A significant amount of research and development has been carried out in order to identify phishing attempts on the basis of the distinctive content, network, and URL characteristics of each attempt. The currently available techniques exhibit great variety with regard to their respective intuitions, data processing procedures, and assessment strategies.

This calls for a thorough systematisation so that the benefits and drawbacks of each technique, as well as their applicability in a variety of settings, can be compared and evaluated in an objective and logical manner. This calls for a thorough systematisation so that the benefits and drawbacks of each technique, as well as their applicability in a variety of settings, can be compared and evaluated in an objective and logical manner.

This article reports on the findings of an extensive study of phishing detection techniques; the emphasis here is on software-based approaches. Starting with the phishing detection taxonomy, studies are undertaken on evaluation datasets, detection features, detection procedures, and evaluation metrics. Finally, we provide some insights that we believe may aid in the creation of more effective and efficient phishing detection techniques. [4]

3. Proposed methodology

A great deal of study has gone into determining if a certain piece of material, network, or URL is indicative of phishing. The available methods have different underlying assumptions, ways of handling data, and methods of evaluation.

This calls for a thorough systematisation that will for a thorough and systematic examination and comparison of the merits and limitations of each technique, as well as their application in different settings. This calls for a thorough systematisation so that the merits and drawbacks of each approach, as well as their suitability in various circumstances, can be investigated and evaluated in a methodical and reasonable manner.

In this post, we'll take a look at some of the methods that may be used to spot phishing attempts. We investigate the datasets, detection properties, detection methodologies, and evaluation metrics proposed by the phishing detection taxonomy. Last but not least, we provide suggestions for improving phishing-detection methods.

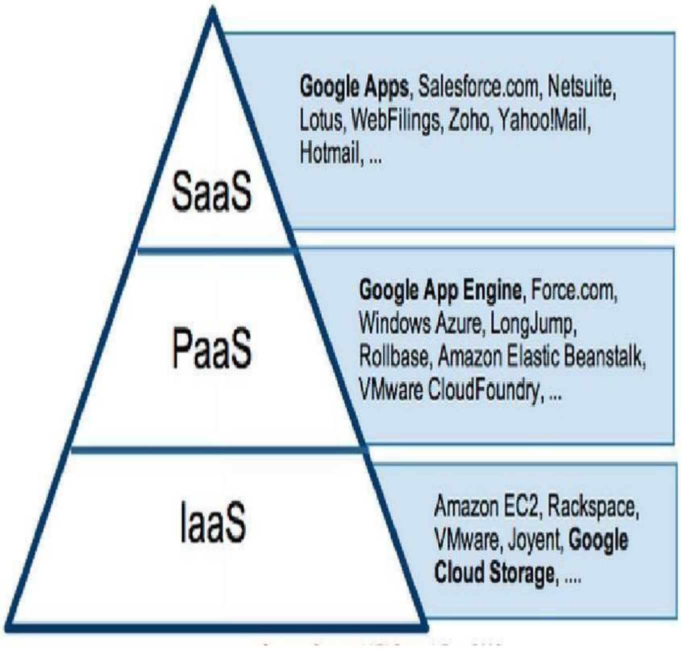

SaaS
Google Apps, Salesforce.com, Netsuite, Lotus, WebFilings, Zoho, Yahoo!Mail, Hotmail, ...
PaaS
Google App Engine, Force.com, Windows Azure, LongJump, Rollbase, Amazon Elastic Beanstalk, VMware CloudFoundry, ...
IaaS
Amazon EC2, Rackspace, VMware, Joyent, Google Cloud Storage, ...

Figure 1: Delivery model in the Cloud.

In Figure 2, we can see examples of different types of clouds. In contrast to private clouds, which provide exclusive use of virtualized resources, public clouds are shared, external cloud environments that several tenants may use. Each community cloud is tailored to a unique set of users.

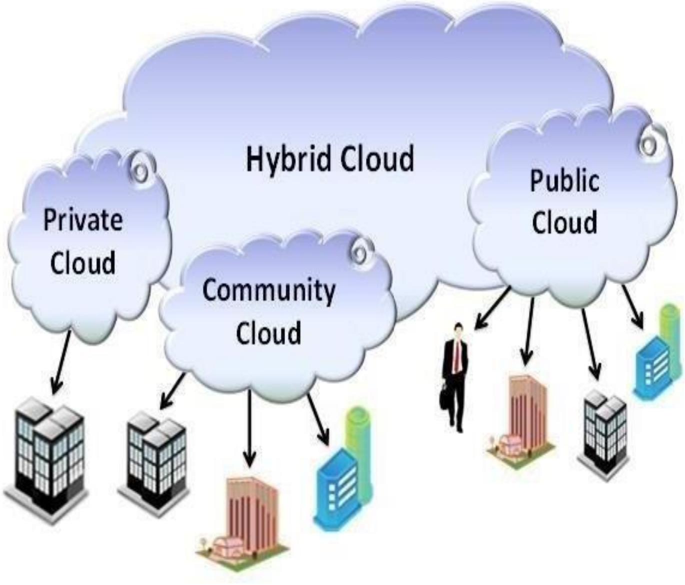


Figure 2: Deployment model in cloud

Listed below is a diagram depicting the proposed system's architecture.

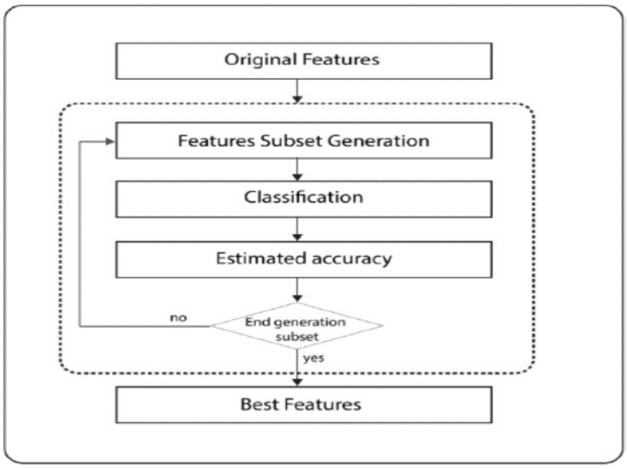


Figure 3: System architecture for the proposed system.

3.1 Data access layer

All interactions with the database are made public through the data access layer. Utils, POJOs, and interfaces for interacting with DAOs are included. The DAO layer is the entry point for data access by all other project modules.

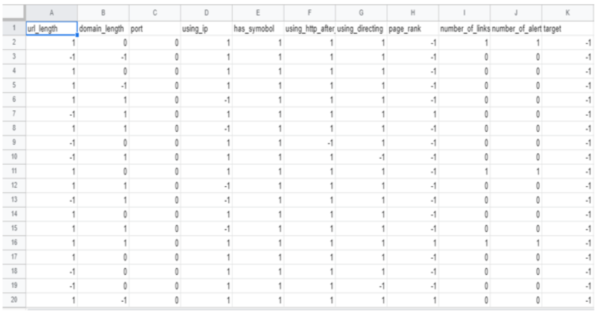
Operations performed in the model is as shown below,

- Create a new vendor or purchaser profile
- The first step is to access your current account by signing in.
- Please logout of this session.
- Changes to the current profile
- Please update your password for security purposes.
- • If you forget your password, you may have the current one sent to you.
- To cancel an existing account:

3.2 Employed dataset

•Often, data from a number of different sources must be combined into one cohesive model. Specifically, three data sets are typically employed during the model development process.

• In the training dataset, each input vector (or scalar) is paired with its associated target output vector (or scalar) (or label). Every input vector in the training dataset is processed by the current model and compared to the goal.



	A	B	C	D	E	F	G	H	I	J	K
1	url_length	domain_length	port	using_ip	has_symobol	using_http_after	using_directing	page_rank	number_of_links	number_of_alert	target
2	1	0	0	1	1	1	1	-1	1	1	-1
3	-1	-1	0	1	1	1	1	-1	0	0	-1
4	1	0	0	1	1	1	1	-1	0	0	-1
5	1	-1	0	1	1	1	1	-1	0	0	-1
6	1	1	0	-1	1	1	1	-1	0	0	-1
7	-1	1	0	1	1	1	1	1	0	0	-1
8	1	1	0	-1	1	1	1	-1	0	0	-1
9	-1	0	0	1	1	-1	1	-1	0	0	-1
10	-1	1	0	1	1	1	-1	-1	0	0	-1
11	1	0	0	1	1	1	1	-1	1	1	-1
12	1	1	0	-1	1	1	1	-1	0	0	-1
13	-1	1	0	-1	1	1	1	-1	0	0	-1
14	1	0	0	1	1	1	1	-1	0	0	-1
15	1	1	0	-1	1	1	1	-1	0	0	-1
16	1	1	0	1	1	1	1	1	1	1	-1
17	1	0	0	1	1	1	1	-1	0	0	-1
18	-1	0	0	1	1	1	1	-1	0	0	-1
19	-1	0	0	1	1	1	-1	-1	0	0	-1
20	1	-1	0	1	1	1	1	1	0	0	-1

Figure 4: Employed dataset

3.3 Training and testing the model

The model will be trained with the help of the datasets, and its performance will be evaluated via testing. It is possible that improvements will be performed to enhance accuracy if the circumstance warrants it. Creating algorithms that can examine data, learn from it, and then use that information to making predictions is a frequent objective in machine learning. Such algorithms are effective because they build a mathematical model from the information they are provided with. The data needed to construct the ultimate model might originate from several places. Specifically, three data sources are included at different stages of the model development procedure.

To begin, a training dataset (a collection of samples used to fine-tune the model's parameters) is utilised for the model's initialization and initialization phases of the training process (such as the weights of connections between neurons in ANNs). Using a supervised learning method (such gradient descent or stochastic gradient descent), the model (which might be a neural network or a naive Bayes classifier) is trained on the training dataset (e.g. gradient descent or stochastic gradient descent). As a practical matter, the training dataset generally consists of pairs of input and output, with the latter being labelled as the target (or label). The current model applies itself to each input vector in the training dataset, and then checks whether or not the generated vector matches the target vector. Variables in the model are fine-tuned in response to the outcomes of the comparison and the chosen learning strategy. The procedure of fitting a model could include either selecting variables to be used or estimating their values as parameters.

Afterwards, the fitted model is used to generate predictions on a second dataset, known as the validation dataset. While experimenting with alternative settings for the model's essential parameters, we may objectively assess the training dataset fit using the validation dataset (such as the number of hidden units in a neural network, for instance) (e.g. the number of hidden units in a neural network). Early halting, which may be accomplished using validation datasets, is a kind of regularisation. overfitting to the training dataset is indicated when the error on the validation dataset increases. As the inaccuracy in the validation dataset shifts, it creates several local minima, which further complicates this fundamental process. As a result, many informal measures have been developed to identify overfitting.

The model fit on the training dataset may be evaluated objectively using the test dataset. A holdout dataset is a test dataset that has never been used for training.**3.4 Cloud deployment process of model**

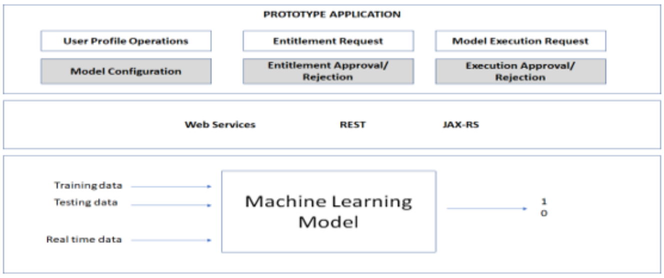


Figure 5: Cloud deployment process on model

3.4.1 Implementation of the model to detect phishing

To determine if a given URL leads to a phishing site, this component uses the machine learning technique Logistic Regression, which is described below.

The logistic function is used to describe a binary dependent variable in the simplest form of the statistical model known as "logistic regression," however there are numerous more complex versions. Logistic regression (sometimes called logit regression) is a kind of regression analysis that uses a logistic model to estimate the model's parameters (a form of binary regression).

3.4.2 Implementation of the model to detect phishing

The following are the features that this module puts into effect: Using a machine learning approach called Logistic Regression, we can determine whether the given URL belongs to a phishing site.

Although many more sophisticated extensions exist, the simplest version of the statistical model known as "logistic regression" employs a logistic function to represent a binary dependent variable. Logistic regression is a kind of regression analysis (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

3.4.3 Implementation of REST services

This module provides an end point to the outside world so that the third party applications can invoke it by sending the new URL. The end point upon receiving the data, will then invoke the seven python programs by sending the data as a command line parameters. This python program will either give output as 1 or 0.

The output 1 indicates it's a phishing site and the output 0 indicates no evidence of phishing found. The end point will then analyse the output from each of the models.

3.4.4 Implementation of Prototype application - User Portal

Functionality for creating and editing user profiles in the prototype application is implemented in this subsystem.

Users may manage their profiles by making new accounts, signing into their current accounts, logging out, modifying their profiles, changing their passwords, and removing their profiles entirely.

The programme is also hosted on the cloud server, making it accessible from anywhere in the world through the server's unique IP address. The J2EE framework is utilised for the implementation, while SQLITE3 is used for the database requirements.

This module covers not only the user profile actions but also additional user-executable operations. Among these activities are asking for what one is entitled to, asking for a forecast, and getting back the results of the requested action. There is no way for the user to carry out the prediction request independently. The administration must approve it first.

3.4.5 Admin portal

This module implements the Admin operations in our prototype application. The admin operations include approving or rejecting the entitlement requests, approving or rejecting the prediction requests, and visualizing the output of the executed requests.

The admin of the project will have the full control on who executes the prediction algorithm and against which data. Any user who wants to run the algorithm against their new data will first have to get it approved from the admin.

The suggested system's block diagram.

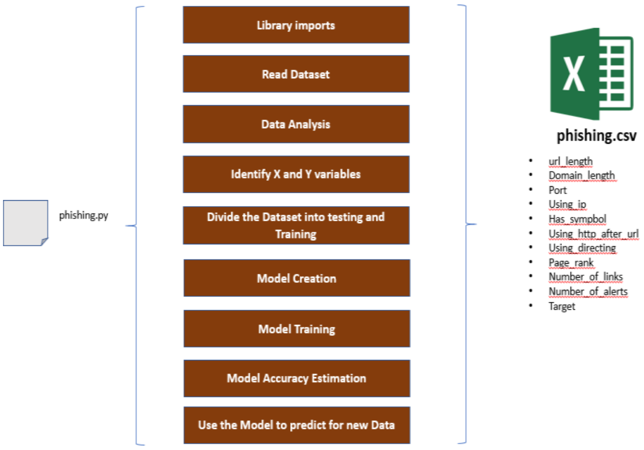


Figure 6: Block diagram for the proposed model.

The following are the features that this module puts into effect: Using a machine learning approach called Logistic Regression, we can determine whether the given URL belongs to a phishing site.

The logistic function is used to describe a binary dependent variable in the simplest form of the statistical model known as "logistic regression," however there are numerous more complex versions. Logistic regression (sometimes called logit regression) is a kind of regression analysis that uses a logistic model to estimate the model's parameters (a form of binary regression).

Researchers in the early 20th century used Logistic Regression in the biomedical sciences. Later applications in the social sciences relied heavily on it. Logistic regression is used in situations where the dependent variable (the objective) is of the categorical kind.

The algorithm for the Logistic regression is briefed in the below figure,

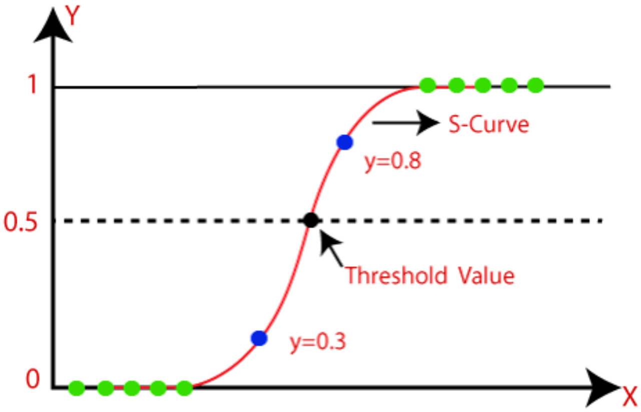


Figure 7: Logistic regression

The above concept has been implemented in the proposed model and it is explained by taking as the example below,

Whether an email is spam (1) or not (2) may be predicted using (0) Indications of whether or not the tumour is cancerous (1) (0) Let's say we're in a position where we need to decide whether or not an incoming email is spam. Using linear regression to solve this issue requires determining a cutoff value for categorization purposes. The real-world repercussions of incorrectly classifying a data item as "not malignant" when the true class is malignant and the anticipated continuous value is 0.4 are severe. This demonstrates why linear regression cannot be used to solve a classification issue. Since linear regression has no limits, logistic regression becomes relevant. Their worth is limited to being exactly 0 or 1.

4. Result and discussion

The home page of the designed website for phishing detection is as shown in the below figure,

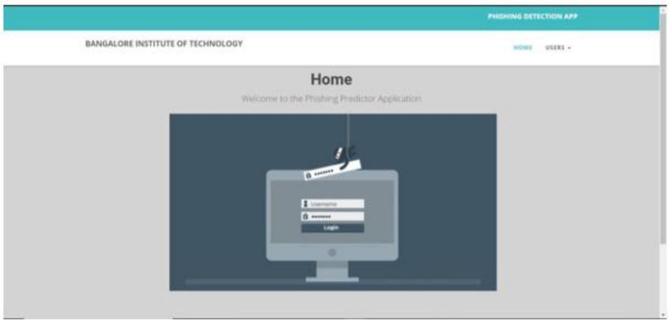

PHISHING DETECTION APP
BANGALORE INSTITUTE OF TECHNOLOGY
HOME
USERS
Home
Welcome to the Phishing Predictor Application
Username
Login

Figure 8: Home page for detection of phishing

The registration page is designed and it allows to register the new user and the same is as shown in the below figure,

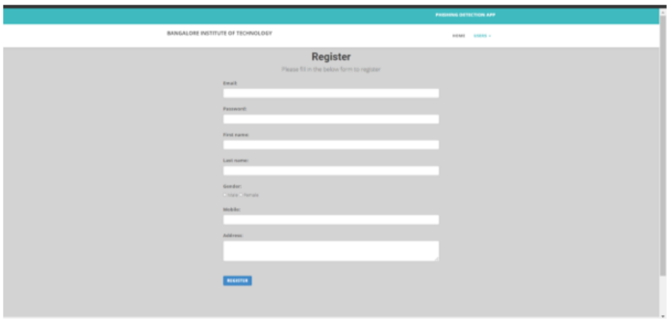


Figure 9: Registration phase for new user.

You may see examples of the administrator and new user login screens in the following image.

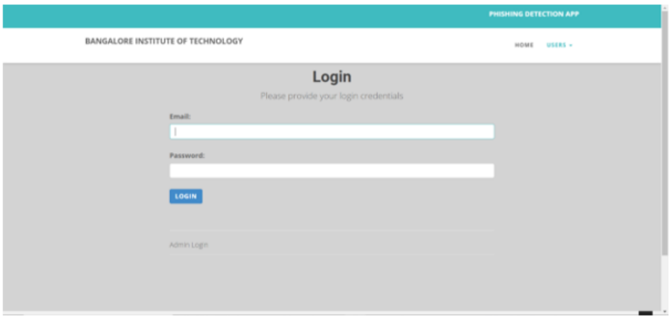

PHISHING DETECTION APP
BANGALORE INSTITUTE OF TECHNOLOGY
HOME
USERS
Login
Please provide your login credentials
Email:
Password:
LOGIN
Admin Login

Figure 10: Login page for the detection website.

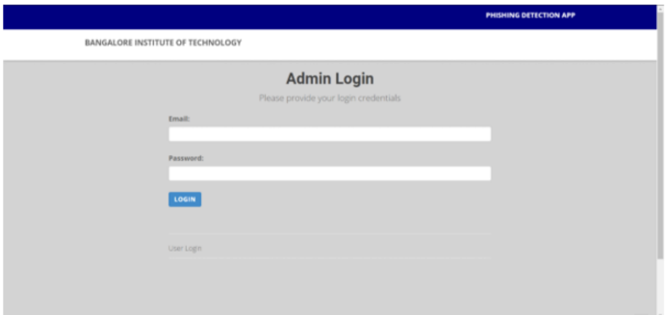
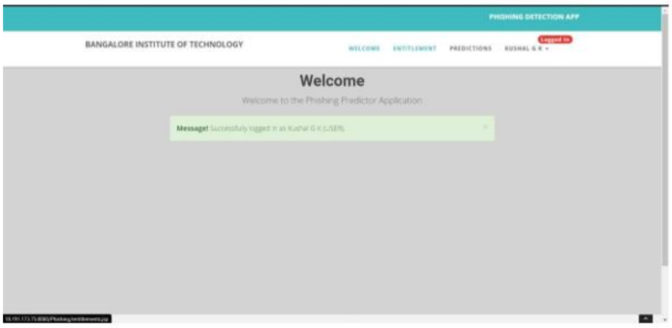

PHISHING DETECTION APP
BANGALORE INSTITUTE OF TECHNOLOGY
Admin Login
Please provide your login credentials
Email:
Password:
LOGIN
User Login

Figure 11: Admin page

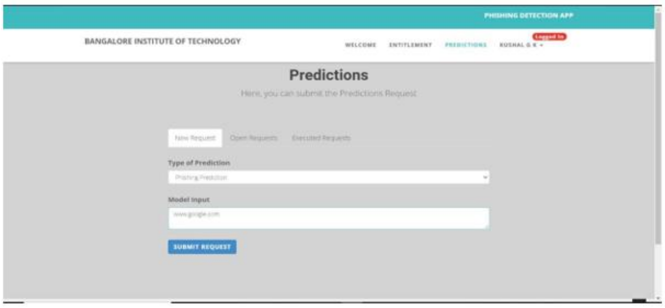
The front page designed is as shown in the below figure after the user gets logged in to the website.


PHISHING DETECTION APP
BANGALORE INSTITUTE OF TECHNOLOGY
WELCOME
ENTITLEMENT
PREDICTIONS
KUSHAL G K
Logged In
Welcome
Welcome to the Phishing Predictor Application

Message! Successfully logged in as Kushal G K (USER)

Figure 12: Front page of the model.

The predication page is designed for the requested URL from the user to check whether link is legitimate or not.

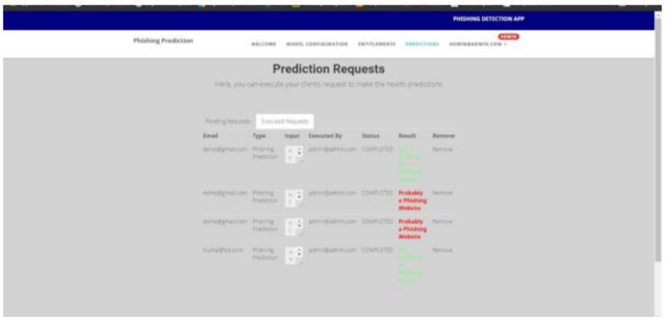


Type of Prediction: Phishing Prediction

Model Input: www.google.com

Figure 13: URL provided by user to check it is legitimate or not.

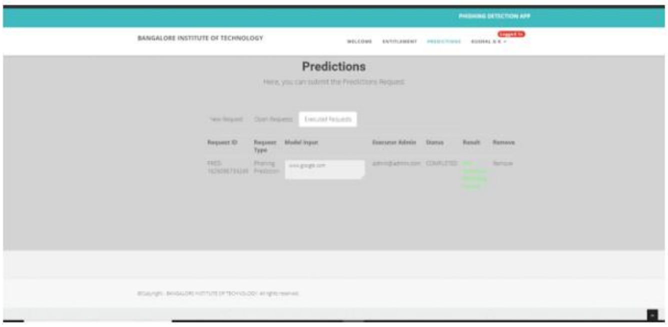
The result provided from the website to check the requirements provided by the user.



Email	Type	Input	Executed By	Status	Result	Remove
[illegible]	Phishing Prediction	[illegible]	admin@admin.com	COMPLETED	[illegible]	Remove
[illegible]	Phishing Prediction	[illegible]	admin@admin.com	COMPLETED	Probably a Phishing Website	Remove
[illegible]	Phishing Prediction	[illegible]	admin@admin.com	COMPLETED	Probably a Phishing Website	Remove
[illegible]	Phishing Prediction	[illegible]	admin@admin.com	COMPLETED	[illegible]	Remove

Figure 14: Predication page for the user requirement.

The legitimate links declared by the model is as shown in the below figure,

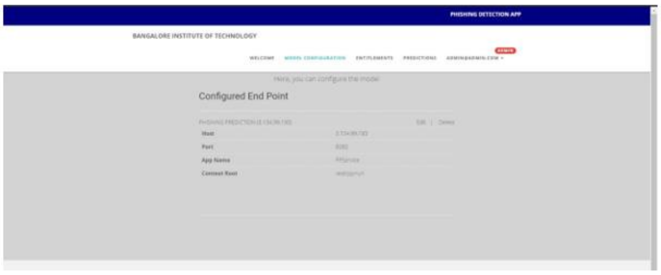


Request ID	Request Type	Model Input	Executor Admin	Status	Result	Remove
PRED-[illegible]	Phishing Prediction	www.google.com	admin@admin.com	COMPLETED	[illegible]	Remove

©Copyright - BANGALORE INSTITUTE OF TECHNOLOGY. All rights reserved.

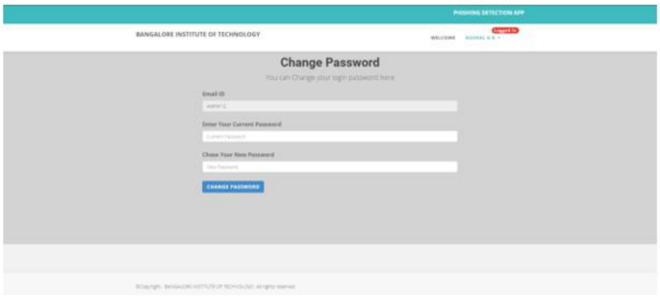
Figure 15: Predication page for the grading the link is legitimate.

The configuration of the page for the model and the password changing option to the user is also provided and the same is as shown in the below figures



PHISHING PREDICTION [illegible]	
Host	[illegible]
Port	[illegible]
App Name	[illegible]
Context Root	[illegible]

Figure 16: Configuration made in the website


PHISHING DETECTION APP

BANGALORE INSTITUTE OF TECHNOLOGY

WELCOME

Logged In

Change Password

You can Change your login password here

Email ID

Enter Your Current Password

Chose Your New Password

CHANGE PASSWORD

©Copyright - BANGALORE INSTITUTE OF TECHNOLOGY. All rights reserved

Figure 17: Password changing page.

5. Conclusion and future enhancement

Here, a trustworthy detection system was developed that is flexible enough to identify phishing pages. Our system use cloud-based machine learning to identify legitimate from malicious internet destinations. The proposed method does not include a server-side or cloud-based service; rather, it relies on the user to gather discriminatory data from website URLs and source code.

This project provides an intelligent phishing website detection system. The technology uses supervised machine learning. We used Logistic Regression for its categorization accuracy. Our goal is to develop a better classifier by researching phishing website attributes.

Future plans include building a browser plugin to employ the suggested method and notify users if they visit a phishing site.

REFERENCES

[1] AO Kaspersky lab. (2017). The Dangers of Phishing: Help employees avoid the lure of cybercrime. [Online] Available: https://go.kaspersky.com/Dangers-Phishing-Landing-Page- Soc.html [Oct 30, 2017].

[2] ”Financial threats in 2016: Every Second Phishing Attack Aims to Steal Your Money” Internet: https://www.kaspersky.com/about/pressreleases/2017 financial-threats-in-2016. Feb 22, 2017 [Oct 30, 2017].

[3] Y. Zhang, J. I. Hong, and L. F. Cranor, ”Cantina: A Content-based Approach to Detecting Phishing Web Sites,” New York, NY, USA, 2007, pp. 639-648.

[4] M. Blasi, ”Techniques for detecting zero day phishing websites.” M.A. thesis, Iowa State University, USA, 2009.

[5] R. S. Rao and S. T. Ali, ”PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach,” Procedia Computer Science, vol. 54, no. Supplement C, pp. 147-156, 2015.

[6] E. Jakobsson, and E. Myers, Phishing and Counter-Measures: Understanding the Increasing Problem of Electronic Identity Theft. Wiley, 2006, pp.2–3.

[7] L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, ”Detecting phishing web sites: A heuristic URL-based approach,” in 2013 International Conference on Advanced Technologies for Communications (ATC 2013), 2013, pp. 597-602.

[8] Z. Zhang, Q. He, and B. Wang, ”A Novel Multi-Layer Heuristic Model for Anti- Phishing,” New York, NY, USA, 2017, p. 21:1-21:6.

[9] N. Sanglerdsinlapachai and A. Rungsawang, ”Web Phishing Detection Using Classifier Ensemble,” New York, NY, USA, 2010, pp. 210-215.

[10] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, ”CANTINA+: A FeatureRich Machine Learning Framework for Detecting Phishing Web Sites,” ACM Trans. Inf. Syst. Secur., vol. 14, no. 2, pp. 21:1-21:28, Sep. 2011.

[11] R. M. Mohammad, F. Thabtah, and L. McCluskey, ”Predicting phishing websites based on self-structuring neural network,” Neural Comput & Applic, vol. 25, no. 2, pp. 443-458, Aug. 2014.

[12] Pradeepthi K V and Kannan A, ”Performance study of classification techniques for phishing URL detection,” in 2014 Sixth International Conference on Advanced Computing (ICoAC), 2014, pp. 135-139.

[13] S. Marchal, J. Franois, R. State, and T. Engel, ”PhishStorm: Detecting Phishing With Streaming Analytics,” IEEE Transactions on Network and Service Management, vol. 11, no. 4, pp. 458-471, Dec. 2014.

[14] Sirageldin, B. B. Baharudin, and L. T. Jung, ”Malicious Web Page Detection: A Machine Learning Approach,” in Advances in Computer Science and its Applications, Springer, Berlin, Heidelberg, 2014, pp. 217-224.

[15] R. Verma and K. Dyer, ”On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers,” New York, NY, USA, 2015, pp. 111-122.

[16] H. H. Nguyen and D. T. Nguyen, ”Machine Learning Based Phishing Web Sites Detection,” in

AETA 2015: Recent Advances in Electrical Engineering and Related Sciences,

[17] V. H. Duy, T. T. Dao, I. Zelinka, H.- S. Choi, and M. Chadli, Eds. Cham: Springer International Publishing, 2016, pp. 123-131.

[18] M. A. U. H. Tahir, S. Asghar, A. Zafar, and S. Gillani, ”A Hybrid Model to Detect 76 Phishing-Sites Using Supervised Learning Algorithms,” in 2016 International Conference on Computational Science and Computational Intelligence (CSCI), 2016, pp. 1126-1133.

[19] M. S. I. Mamun, M. A. Rathore, A. H. Lashkari, N. Stakhanova, and A. A. Ghorbani, ”Detecting Malicious URLs Using Lexical Analysis,” in Network and System Security: 10th International Conference, NSS 2016, Taipei, Taiwan, September 28-30, 2016, Proceedings, J. Chen, V. Piuri, C. Su, and M. Yung, Eds. Cham: Springer International Publishing, 2016, pp. 467-482