IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF PHISHING WEBSITE FOR DATA SECURITY USING VARIOUS MACHINE LEARNING ALGORITHMS

[1]P. Vimala Manohara Ruth, [2]Dr. Y. Rama Devi, [3]E. Haritha, [4]N. Shiva Kumar

[1]Assistant Professor, [2]Professor, [3]Bachelor of Engineering Student, [4]Bachelor of Engineering Student
[1]Department of Computer Science and Engineering,
[1]Chaitanya Bharathi Institute of Technology, Gandipet, Hyderabad, India

Abstract: Phishing is one in all the foremost threats during this net era. Phishing may be a sensible method wherever a legitimate web site is cloned and victim's area unit lured to the pretend web site to supply their personal yet as counselling, typically it proves to be expensive. Although most of the websites can provide a disclaimer warning to the users concerning phishing, users tend to neglect it. It is not a totally accountable action by the websites additionally and there is not a lot of that the websites might very do concerning it. Since phishing has been in persistence for an extended time, several approaches are projected in past which will find phishing websites however only a few or none of them find the target websites for these phishing attacks, accurately.

Our projected methodology tends to establish phishing websites employing a combined approach by constructing Resource Description Framework (RDF) models and mistreatment ensemble learning algorithms for the classification of internet sites. Our approach uses supervised learning techniques to coach our system. As our system explores the strength of RDF and ensemble learning ways and each these approaches work hand in hand, an extremely promising accuracy rate of 97% is achieved.

***Index Terms* – Phishing, Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, XGBoost Classifier**

Introduction

Phishing attack is a typical way of attack or fraud in which an attacker tries to learn sensitive information or data such as login or sign in credentials or account information by sending as a well-known entity or person in email or other communication channels. Typically, a victim or person receives an email or a message that appears to possess been sent by a known contact or organization. The message contains harmful software targeting the user's computer or has links to direct victims to malicious websites in order to trick them into communicating personal and financial information, such as passwords, account IDs or master-card details. Phishing is popular among attackers, since it's very easier to trick someone into clicking a malicious link which seems legitimate than trying to interrupt through a computer's defense systems. These malicious links that are present within the body of the message are designed to make it appear that they go to the send up organization using that organization's logos and other legitimate contents. Many users without their knowledge click these phishing domain links every day and every hour. The attackers are targeting both the companies and the users. According to the MCSIR, released in February 2014, the annual worldwide impact of phishing could be very high as $5 billion. The main reason behind this is the lack of awareness of users. Security defenders must be responsible to take precautions to prevent users from confronting these harmful sites. Preventing these type of attack huge costs can start with making people conscious in addition to build strong security mechanisms which are able to detect and prevent these phishing domains from reaching the user. One of the common victims of these attacks are users of social media network sites. With the huge rise in the internet, the risk associated with phishing attack will also increase and criminals will always try to find new ways to deceive people. As phishing evolves, stakeholders will always find new ways of mitigating the risk associated with phishing.

1.1 Problem Definition

Develop a phishing website detection system through which a user will confirm whether or not an address might be a phished, suspicious or legitimate one. The choices of the URL are extracted and examined in classifying the online site supported the results obtained. A system that provides data and insight to the inexperienced internet users in characteristic the phishing URLs and offers associate information concerning the online site options that are utilized in predicting the phishing attacks. It also evaluates the classification accuracy of a phishing dataset victimization Confusion Matrix.

1.2 Methodology

This paper attempts to evaluate different machine learning techniques that aims to investigate the potential uses of some ensemble classification models in detecting phishing websites. In particular, the aim here is development of an ensemble model that will be used for predicting whether a website is phishing or legitimate, and if so to what degree. At this stage determining the phishing website can be viewed as a data mining classification problem, wherein this instance the class attribute is the degree of phishing. The classification process is based upon attributes and characteristics which are used to distinguish phishing sites such as spelling mistakes, long URL's, prefixes and suffixes. These attributes are obtained from input website URL's. Therefore, an intelligent ensemble learning model to predict phishing websites has to be designed and developed. After the prediction of URL, based on the features extracted, the webpage is displayed to the end user informing if the website is either phishing or not.

I. Literature review

The methods of detecting phishing attacks are:
Google Safe Browsing: This approach uses the blacklist URLs to discover the phishing attack. A sample URL is taken as input and checked within the blacklist repository. If the URL is present in the black list repository, the URL is termed as suspicious URL, else it is a legitimate website. The main shortcoming of this approach is its inability to detect the phishing URL which aren't present within the blacklist which could increase the false positives rate.
Spoof Guard: This method scans suspicious websites for phishing symptoms to determine whether the website is legitimate or phishing. Some heuristics include image verification, link verification, URL verification and password field verification. If the total score of the phishing symptoms listed above exceeds the threshold, it is classified as a legitimate phishing website. This method detects zero-day attacks. This method also has a high limit on the number of false positives.
False alert: This method will use visual phishing detection when the attacker uses the same CSS style to deceive the original website. In this method, CSS style comparisons are performed on whitelisted websites with suspicious website styles to detect phishing. attack.

Table 2.1: Existing Methods and their disadvantages

Method	**Disadvantages**
Early detection and manual blocking of phishing sites.	Most Internet users do not know how to identify phishing websites in real time. Even experienced people are often attacked for forgetting to check the legitimacy of the website. They do not provide safety training when they busy at work.
Detection of website content and URL [4]	The URL detection of new website is insufficient. These methods are not precise and usually produce a small number of false positives.
Block the phishing E-Mails by various spam filter software [3]	These spam filters tend to block genuine messages. They fail to find these attacks excluding from email-threads.
Server – side Detection	Users can receive delayed responses from servers concerning the credibility of the web site. They underperform in slow internet connections.
Client – side Detection	These software's signature - based security controls are proving less and less effective as years pass by. For example, these solutions are not particularly good at identifying file – less malware. They utilize a lot of memory.
Other Detection Methods	It is not effective on pages that are not visited previously and websites should be maintained by constantly updating to preserve better accuracy.

Some researchers use the universal resource locator to compared them with existing blacklists that contain lists of malicious and phishing websites and their URLs, they have been creating, associated to this there are others those who have used the URL in an opposite manner, particularly comparing the URL with a whitelist of legitimate websites. The latter approach uses heuristics algorithm, that uses a particular database of any known attacks that match the signature of the heuristic pattern to decide if it is a phishing website. Additionally, measuring website traffic by using Alexa is completely different way that has been implemented by researchers to detect phishing websites.

Dilbag Singh, Saloni Manhas, and Swapnesh Taterh "A Novel based Approach for Phishing Websites Detection using Decision Tree" [1]
By making use of decision tree algorithm to classify information gain, financial gain and other uncertainty FST to increase the performance of anti-phishing detection application. In order to overcome issues of phishing attack, anti-phishing detection was designed to detect phishing website URLs on the victim's email. Besides that, anti-phishing detection application able to generate a report of phishing website which are attached on victim's email.

Jitendra Kumar, A. Santhanavijayan, B. Janet, B.S. Bindhumadhava, and Balaji Rajendran “Phishing Website Classification and Detection Using Machine Learning” [2]
By making use of lexical structure od URL to classify url into different parts and identify the Url whether the given url is phishing url or not. In, this paper, they have compared different machine learning techniques for the phishing URL classification task and achieved the highest accuracy of 96% for Naïve Bayes Classifier with a precision=1, recall = .95 and F1-Score= .96.

Chua Shang Ren; Rabab Alayham Abbas Helmi; Muhammad Irsyad Abdullah; Arshad Jamal “Email Anti-Phishing Detection Application” [3]
There are many techniques to overcome tricked by phishing website. One of the methods mostly used to detect phishing is by using visual similarity. This method is to dissimilar phishing webpage, which also reduce the successful rate of victim got tricked by phishing scams. Besides that, there is another method to detect phishing website is by using compression algorithm. Compression algorithm is a critical component which perform a compression of nine compressors that include 1-dimensional string and 2-dimensional image compression.
The main aim of this paper is to spot phishing attacks that connects the victim’s email by mistreatment by applying decision tree algorithm that enforced within the application. This project mainly focused on detect on attachment file of phishing website in the email buddies by using decision tree algorithm. Anti-Phishing detection application used to detect, identify and block the phishing website or email that effected by the phishing website. It is able to calculate the percentages of stored phishing emails in the user’s email.

Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, Aram Alsedrani “Detecting Phishing Websites Using Machine Learning” [4]
The system acts as an extra functionality to a web browser as extension that mechanically notifies the user once it detects a phishing web site. The system is predicated on a machine learning method, notably supervised learning. They've got selected the Random Forest technique because of its sensible performance in classification.
The focus will be on the features combination that we get from Random Forest (RF) technique, as it has good accuracy, is relatively robust, and has a good performance.
Recently, there have been several studies that are trying to solve the phishing problem. They can be classified into four types: blacklist, heuristic, content analysis, and machine learning techniques. The blacklisting technique compares the URL with an existing database that contains a list of phishing website URLs. Because of the rapid increase of such phishing attacks, the blacklist approach has become more inefficient in checking whether each URL is a phishing website or not, and this kind of delay can also lead to zero-day attacks from these new phishing sites.

Fuma Dobashi, Akihito Nakamura, “Proactive Phishing Sites Detection,”[5]
In this paper, emphasized compared phishing mitigation techniques, such as blacklist, heuristics, visual similarity, and machine learning and concluded that these techniques have limitations in dealing with zero-hour attacks and proactive detection of phishing websites. The authors proposed suspicious URL’s generation and to predicts likely phishing sites from the given legitimate brand domain name and scores and judges suspects by calculating various indexes to detect phishing websites.

Akbar-Siami Namin, Moitrayee Chatterjee, “Detecting Phishing Websites through Deep Reinforcement Learning”[6]
This paper proposed a deep reinforcement learning model to detect malicious URLs. This model is capable of adapting to the dynamic behavior of the phishing sites and thus even learn the features associated with phishing website detection. The proposed model uses Deep Reinforcement Learning Techniques. They got an accuracy of 90%.

Dimitris Tsaptsinos; Martyn Weedon; James Denholm-Price “Random Forest explorations for URL classification”[7]
This paper builds the classifier using Random Forest techniques. These RF techniques are used to classify the given url into substring and then consider them. In this paper, the main objective is to evaluate the performance of the Random Forest algorithm using a lexical only dataset. The performance is benchmarked against some other machine learning techniques and additionally against those reported in the literature. Initial results from experiments indicate that the Random Forest algorithm performs the best yielding an 86.9% accuracy.

Table 2.2: Literature Survey of various papers

AUTHOR	YEAR	TITLE	FINDINGS
Saloni Manhas, Swapnesh Taterh Dilbag Singh [1]	2020	A Novel based Approach for Phishing Websites Detection using Decision Tree	A Novel Approach for Phishing Websites Detection was proposed. This model allows to detect weather given website is phishing website or not using Decision Tree Algorithm.
Chua Shang Ren; Rabab Alayham Abbas Helmi; Arshad Jamal; Muhammad Irsyad Abdullah [3]	2019	Anti-Phishing Email Detection Application	This paper describes how to detect Phishing Websites that are sent to our Email. This model is being developed by Agile Unified Process.
Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, Aram Alsedrani [4]	2019	Detecting Phishing Websites Using Machine Learning	This model describes how to identify Phishing Websites using Random Forest Algorithm.

Martyn Weedon; Dimitris Tsaptsinos; James Denholm-Price [7]	2017	Random forest explorations for URL classification	This model describes classification of URL's using Random Forest Algorithm.

II. Design of the Proposed System

The system design consists of three main phases, namely feature selection, modelling, and evaluation. The module is implemented to extract the features from the input site. In the proposed model illustrate the association rule mining algorithms on a phishing URL data set, machine learning repository. This phase aims to select to the most significant features such as text, URL, log data, and more to distinguish between legitimate and phishing websites. While in the modelling phase, a ensemble framework will be developed to handle the selected most significant features from the first phase. Different algorithms will be applied which is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest.

The final phase is the evaluation phase which aims to assess the overall performance of the suggested classification framework, therefore, the most widely applied evaluation metrics for phishing detection problems such as classification accuracy, g-mean, F1 evaluation, precision, recall will be applied in this phase.

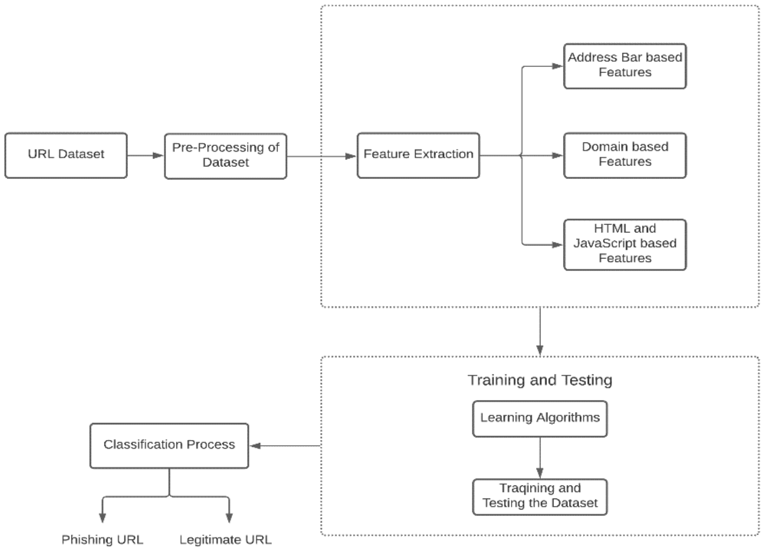


Fig 3.1: Block Diagram of the Proposed System

The above figure 3.1 block diagram describes the steps in constructing a model for Phishing Website Detection. As shown figure 3.1, first data is retrieved from datasets. The data has to prepared for training. So, the data is undergone into pre-processing step for feature extraction or feature selection. Now, the model is constructed with machine learning algorithms with the retrieved features. Now, the model is trained with dataset. The model is validated and tuned to improve the classification accuracy of model. Finally, the model is tested with test dataset to know whether the model is predicting and detecting whether given url is phishing or not.

3.1 Algorithms

Machine learning provides optimized and efficient ways for data and knowledge analysis. Recently, it has shown encouraging results in real-time classification problems. One of the main benefits of machine learning is the ability to create flexible models for specific tasks, such as phishing detection. Machine learning models can be used as powerful tools. Machine learning models can quickly adapt to changes to identify fraudulent transaction patterns, thereby helping to design learning-based identity systems. Most of the machine learning techniques mentioned here are classified as the supervised machine learning. Here, the algorithm tries to learn a function that compares input and output based on the pattern of input and output pairs. Derive a function from labelled training data with a series of training examples. Used in our research.

3.1.1 Logistic Regression

Logistic Regression is a one of the class set of classification algorithm that is used to classify observations to a series of discrete classes. Logistic regression different from linear regression that generates outputs of continuous number values, whereas Logistic Regression uses logistic sigmoid function to convert its output to return a probability value that can then be mapped to two or more set of discrete classes. Logistic regression is well suitable when the relationship of the data is almost linear despite if there are complex nonlinear relationships between different variables, and it has poor performance. Along with these, it also requires more statistical assumptions before using other techniques.

Logistic Regression is a statistical method that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or Logit Regression) is estimating the parameters of a logistic model.

3.1.2 Support Vector Machine

The idea behind SVM is to use the maximum distance between classes to find the closest point between two classes. This method is a supervised learning model used for linear and nonlinear classification[7]. Nonlinear classifications performed using a kernel function to map the input to a higher-dimensional feature space. Although SVMs are very effective and are very widely used in classification, they also have some disadvantages. These requires high calculations to train the data. Also, data is prone to over-fitting. The four common kernel functions at the SVM are Radial basis function, linear function, polynomial and sigmoid.

3.1.3 Decision Tree

The decision tree classifier is used as a well-known classification method. A decision tree is a tree structure similar to a block diagram, in which internal nodes represent elements or attributes, branches represent decision rules, and each leaf node represents a result. The node in the decision tree is called the root node. Learn to block by attribute value. Split the tree recursively, which is called recursive splitting. This feature provides higher resolution for processing large numbers of digital or categorical data sets. Decision trees are also ideal for dealing with non-linear relationships between attributes and classes. The fouling function is determined periodically to evaluate the separation quality of each node, and the Gini coefficient is used. In fact, decision trees are flexible in the sense that they can easily model nonlinear or non-standard relationships. ferry. You can explain the interaction between the predictor variables. Due to its binary structure, it can also be explained well. However, decision trees have some shortcomings. These shortcomings tend to overuse data, and it is difficult to update the decision tree with new choices [1][2].

3.1.4 Random Forest

As the name suggests, a random forest contains a large number of individual decision trees, which are used as a group to make initial decisions. Each tree in a random forest defines a prediction category, and the result is the most predictable category of causes. Because the surprising result of accidental forests is that trees can protect each other from individual mistakes. Although some trees may predict the wrong answer, many others correct the final prediction. As a whole, trees can move around in a random forest. By combining many weak learners who cannot adapt because only a subset of A is used, overfitting can be reduced. In all training examples, random forest can handle a large number of variables in the data set. In addition, they made an objective assessment of the generalization errors in the afforestation process. They can also make a good estimate of the missing data. The main disadvantage of random scaffolding is the lack of repeatability, because the scaffolding process is random. In addition, since it contains many independent decision trees, it is difficult to interpret the final model and subsequent results [3][4].

3.1.5 XGBoost Classifier

Recent research has found that this algorithm, its methodology and its application in machine learning classification are very useful. It is very fast and has better performance because it is a managed decision tree execution. The classification model is used to improve model performance and speed.

3.2 Feature Selection

3.2.1 Address Bar based Features [4]:

1. Using IP Address: If IP address is used in place of the name of the domain in the URL, for example, “http://125.98.3.123/fake.html”, users can be sure that someone are trying to steal user’s personal information. In most times, an IP address can be transformed to the hexadecimal code as shown in the following link “http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html”.

$$\text{Rule: IF}\begin{cases}\text{If Domain name contains IP Address } \rightarrow \text{ Phishing}\\ \qquad\qquad\text{Otherwise} \rightarrow \text{ Legitimate}\end{cases}$$

2. Using the URL shortening Services “Tiny URL”: URL shortening is a technology on the “World Wide Web” in which a URLs can be smaller in length and still point to the desired webpage. This is achieved by means of an “HTTP Redirect” to a short domain name, which results in a web page with a long URL.

$$\text{Rule: IF}\begin{cases}\text{ TinyURL } \rightarrow \text{ Phishing}\\ \text{Otherwise} \rightarrow \text{ Legitimate}\end{cases}$$

3. URL with “@” Character: When using the "@" character in the URL, the browser ignores anything before the "@" character, and the actual address is usually after the "@" character.

$$\text{Rule: IF}\begin{cases}\text{Url Having @ Symbol} \rightarrow \text{ Phishing}\\ \qquad\text{Otherwise} \rightarrow \text{ Legitimate}\end{cases}$$

4. Redirecting with "//": The presence of" //" in the URL path means that the user has been redirected to another website. We check where "//" appears. If the URL starts with "HTTP", it means that "//" should be in the sixth position. However, if the URL uses "HTTPS", "//" should be in the seventh place.

$$\text{Rule: IF}\begin{cases}\text{The Last Occurrence of "//" in the URL } > \ 7 \rightarrow \text{ Phishing}\\ \qquad\qquad\text{Otherwise} \rightarrow \text{ Legitimate}\end{cases}$$

5. Add a prefix or suffix to your domain and separate it with (-): hyphens are rarely used in valid URLs. Subscribers usually add prefixes or suffixes separated by (-) to domain names to make users feel that they are dealing with legitimate domain names. website. For example, http://www.Confirme-paypal.com/

$$\text{Rule: IF}\begin{cases}\text{Domain Name Part Includes } (-) \text{ Symbol } \rightarrow \text{ Phishing}\\ \qquad\qquad\text{Otherwise } \rightarrow \text{ Legitimate}\end{cases}$$

6. Domain name registration time: As phishing websites exist for a short time; we believe that trusted domain names will pay regularly several years in advance. In our data set, we found that for a long time, the most fraudulent domain name has been used for the one-year

$$\text{Rule: IF}\begin{cases}\text{Domains Expires on} \le \ 1 \text{ years } \rightarrow \text{ Phishing} \\ \qquad\qquad \text{Otherwise} \rightarrow \text{ Legitimate}\end{cases}$$

3.2.2 Domain based Features:

1. Domain age: This feature can be obtained from a database called WHOIS. Most phishing sites do not have a long lifespan. Looking at our data set, we found that the minimum age for a legal domain name is 6 months

$$\text{Rule: IF}\begin{cases}\text{Age Of Domain} \ge 6 \text{ months } \rightarrow \text{ Legitimate} \\ \qquad\qquad \text{Otherwise } \rightarrow \text{ Phishing}\end{cases}$$

2. DNS records-For phishing websites, the WHOIS database cannot identify the claimed identity or the host name record cannot be found. If the DNS record is empty or not found, the website is classified as a spoof. Otherwise, it can be called as a "legitimate"

$$\text{Rule: IF}\begin{cases}\text{no DNS Record For The Domain } \rightarrow \text{ Phishing} \\ \qquad\qquad \text{Otherwise} \rightarrow \text{ Legitimate}\end{cases}$$

3. Website Traffic: This feature is used to measures the popularity of the website by identifying the number of visitors and the number of pages they visit. However, because phishing websites only exist for a short time, the Alexa database may not be able to identify phishing websites (Alexa Web Information Corporation, 1996). After checking our data set, we found that the top ranking of legitimate websites is 100,000. If the domain has no traffic or the Alexa database cannot identify it, it is classified as "phishing". Otherwise, it will be classified as suspicious.

$$\text{Rule: IF}\begin{cases}\text{Website Rank} < 100{,}000 \ \rightarrow \text{ Legitimate} \\ \text{Website Rank} > 100{,}000 \ \rightarrow \text{Suspicious} \\ \qquad\qquad \text{Otherwise } \rightarrow \text{ Phish}\end{cases}$$

3.2.3 HTML and JavaScript based Features:

1. Site redirection. The thin line that distinguishes a phishing website from a legitimate website is the number of website redirects. In our data set, we found that legitimate websites redirect once at most. Redirected at least 4 times.

$$\text{Rule: IF}\begin{cases}\qquad \text{ofRedirect Page} \le 1 \ \rightarrow \text{ Legitimate} \\ \text{of Redirect Page} \ge 2 \text{ And } < 4 \rightarrow \text{ Suspicious} \\ \qquad\qquad \text{Otherwise } \rightarrow \text{ Phishing}\end{cases}$$

2. Iframe redirection: IFrame is an HTML tag used to display other web pages on top of the currently displayed web page. Phishers can use iframe tags and hide them. I. There is no frame. With this in mind, phishers use the frameborder attribute, which forces the browser to display a visual outline.

$$\text{Rule: IF}\begin{cases}\text{Using iframe} \rightarrow \text{ Phishing} \\ \text{Otherwise } \rightarrow \text{ Legitimate}\end{cases}$$

III. IMPLEMENTATION

The steps for implementation are:

Step 1: Collect dataset containing phishing and legitimate websites from the open-source platforms.

Step 2: Write a code to extract the required features from the URL database.

Step 3: Analyse and pre-process the dataset by using EDA techniques.

Step 4: Divide the dataset into training and testing sets.

Step 5: Run selected machine learning algorithms like SVM, Random Forest, logistic regression, decision tree and XG Boost on the dataset.

Step 6: Write a code for displaying the evaluation result considering accuracy metrics.

Step 7: Compare the obtained results for trained models and specify which is better.

4.1 FLOWCHARTS/DFD’S/ ER DIAGRAMS

A flowchart is simply a graphical representation of steps. It shows steps in sequential order and is widely used in presenting the flow of algorithms, workflow or processes.

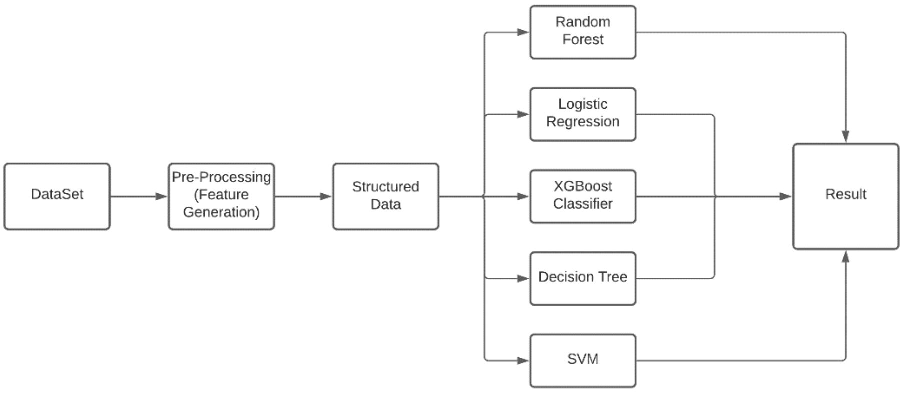


Fig 4.1: Flowchart of the Phishing Website Detection

4.3 Data Set Description

The data set used is downloaded from Kaggle Repository. It contains around 10,000 URLS along with some features. Out of these, 5000 are phishing website URL’s and 5000 are legitimate website URL’s. The data set also contains 17 features that are extracted from each URL. The attribute provides information such as the having IP address, age of the domain, URL length, URL depth, redirection, iframe, etc. Each feature value holds some categorical value, either binary or ternary. Binary value indicates that the existence or the lack of existence of the feature within the URL determines the value assigned to the feature. For ternary features, the existence of the feature in a specific ratio determines the value assigned to the feature. The features that we used in this research work are described in the following.

1. Having IP address in the URL: The usage of an IP address in the domain name is an indicator of a non-legitimate website.
2. Web Traffic: High web traffic indicates that website is used regularly and is likely to be legitimate.
3. URL length: Phishing websites often use long URLs so that they can hide the suspicious part of the URL.
4. Age of the domain: Domain that are in service for a longer period of time are likely to be legitimate.
5. Popup Windows: Usually, legitimate sites do not ask users their credentials via popup windows.
6. Request URL: Often, in legitimate websites, objects are loaded from the same domain where the webpage is loaded.

IV. RESULTS AND DISCUSSION

The code was implemented using Jupyter notebook with the machine learning algorithms using Python Language. A comparative analysis was performed to predict the phishing website URL’s using the following algorithms: Random Forest, XG Boost, Logistic Regression, Support Vector Machine and Decision Tree. Based on the accuracy generated by these algorithms, the best algorithm is considered for predicting whether the URL is Phishing or Legitimate.

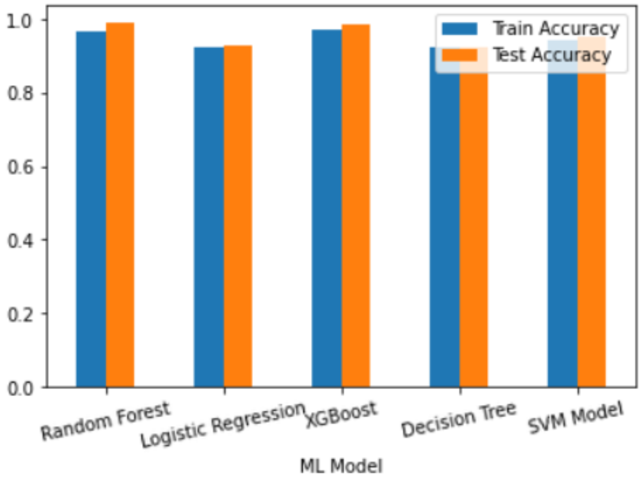
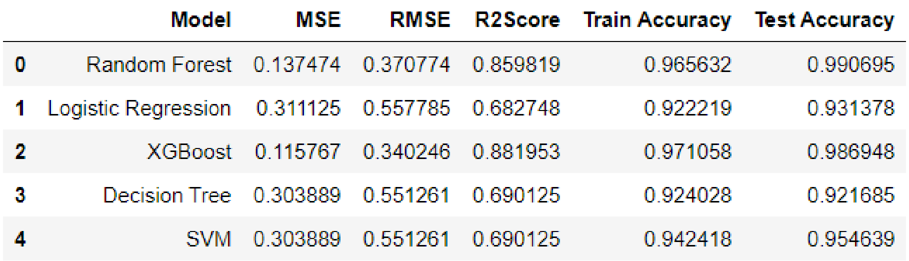


Fig 5.1: Training and Testing Accuracy for different models

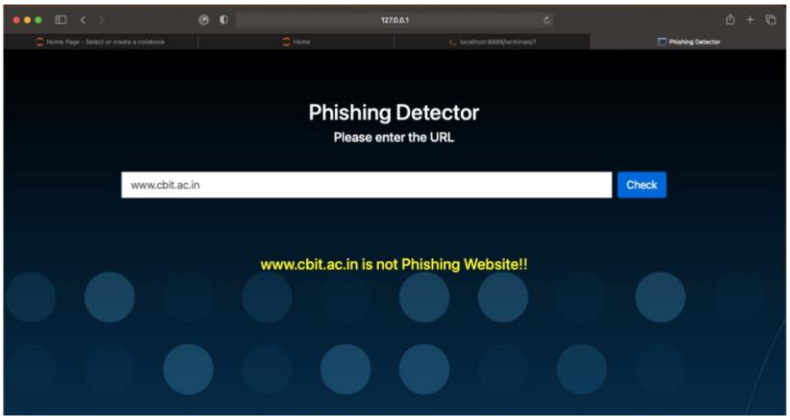


	Model	MSE	RMSE	R2Score	Train Accuracy	Test Accuracy
0	Random Forest	0.137474	0.370774	0.859819	0.965632	0.990695
1	Logistic Regression	0.311125	0.557785	0.682748	0.922219	0.931378
2	XGBoost	0.115767	0.340246	0.881953	0.971058	0.986948
3	Decision Tree	0.303889	0.551261	0.690125	0.924028	0.921685
4	SVM	0.303889	0.551261	0.690125	0.942418	0.954639

Fig 5.2: Comparision of Different Models Accuracies

Figure 5.1 and Figure 5.2 shows the accuracy of different models that have been trained to predict the phishing website URL’s. The highest accuracy was gained by the Random Forest Model and then another best algorithm to gain the accuracy was XGBoost

Classifier algorithm. Random Forest Algorithm was able to gain the accuracy of 97% for training the model and 98% for testing the model.


127.0.0.1
Home Page - Select or create a notebook
Home
localhost:8888/terminals/1
Phishing Detector

Phishing Detector

Please enter the URL

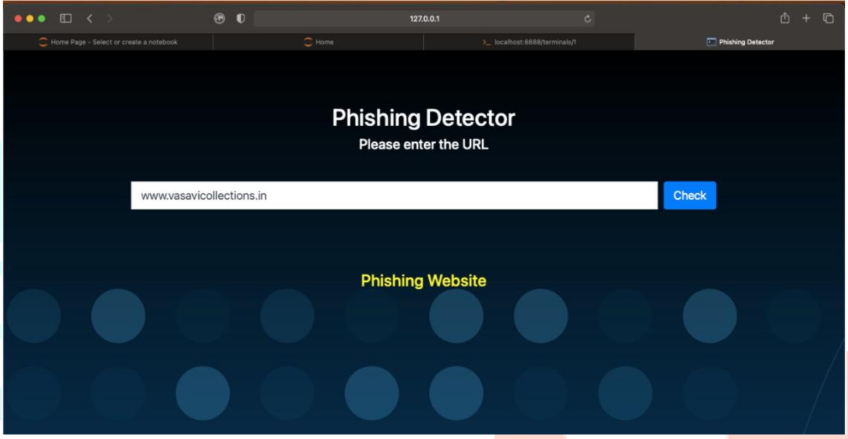
www.cbit.ac.in

Check

www.cbit.ac.in is not Phishing Website!!

Fig 5.3: Phishing Detector output for legitimate website URL

Figure 5.2 shows the predictions of a legitimate URL. This is the prediction made by 17 features extracted from the dataset. This is the GUI for the end user to give a URL to predict whether given website URL is phishing website or legitimate website.


127.0.0.1
Home Page - Select or create a notebook
Home
localhost:8888/terminals/1
Phishing Detector

Phishing Detector

Please enter the URL

www.vasavicollections.in

Check

Phishing Website

Fig 5.3: Phishing Detector output for phishing website URL

Figure 5.3 shows the prediction of Phishing URL. The model learns from the pattern of training data and then predict the URL. Phishing website can be mainly classified and predicted based on the domain-based features such as age of the domain. Sometimes, only age of the domain can only predict the phishing URLs.

V. Conclusion and Future Scope

It has been observed that phishing poses a major threat to network security, and phishing detection is a major problem. We have studied some traditional phishing detection methods. Namely blacklist and heuristic evaluation methods and their shortcomings. According to various characteristics of URLs, a lexical analysis model is created to detect phishing or non-phishing websites in Python, and calculate the accuracy of various algorithms. These methods can identify phishing websites and mitigate social engineering attacks. The project can be further developed by creating a browser extension for GUI development and more features can be considered for classification of the phishing websites.

References

[1] Dilbag Singh, Saloni Manhas, and Swapnesh Taterh "A Novel based Approach for Phishing Websites Detection using Decision Tree" International Journal of Advanced Science and Technology, 29(3), 2020

[2] Jitendra Kumar, A. Santhanavijayan, B. Janet, B.S. Bindhumadhava, and Balaji Rajendran "Phishing Website Classification and Detection Using Machine Learning" - 2020 International Conference on Computer Communication and Informatics

[3] Chua Shang Ren; Rabab Alayham Abbas Helmi; Muhammad Irsyad Abdullah; Arshad Jamal "Email Anti-Phishing Detection Application" - IEEE International Conference on System Engineering and Technology, 2019

[4] Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, Aram Alsedrani "Detecting Phishing Website Using ML" – ICCAIS (International Conference on Computer Application & Information Security), IEEE, 2019

[5] Akihito Nakamura, Fuma Dobashi, "Proactive Phishing Sites Detection," IEEE/WIC/ACM International Conference on Web Intelligence), pp. 443-448, October 2019

[6] Moitrayee Chatterjee, Akbar-Siami Namin, "Detecting Phishing Websites through Deep Reinforcement Learning," IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), July 2019

[7] Martyn Weedon; Dimitris Tsaptsinos; James Denholm-Price "Random Forest explorations for URL classification" – 2017 IEEE International Conference on Cyber Situational Awareness, Data Analytics and Assessment (Cyber SA)