Detection of Phishing Website Using ML

D K Shwetha[1], Pooja N Kulkarni[2], Sahana S Nayak[3], Rajendra M[4]
[1, 2, 3, 4] Dept of CS&E
[1, 2, 3, 4] Atria Institute of Technology, Bangalore

Abstract- *Phishing website is one in every of the online security problems that specialize in the human vulnerabilities rather than software vulnerabilities. it's described because the method of attracting online users to induce their sensitive information like usernames and passwords. during this paper, we offer an intelligent system for detecting phishing websites. The system acts as an additional functionality to an online browser as an extension that automatically notifies the user when it detects a phishing website. The system relies on a machine learning method, particularly supervised learning. we have got selected the Random Forest technique thanks to its good performance in classification. Our focus is to pursue the following performance classifier by studying the features of phishing website and choose the upper combination of them to educate the classifier. As a result, we conclude our paper with accuracy of 98.8% and combination of 26features.*

Keywords- Phishing websites, Random Forest Technique, ML

I. INTRODUCTION

The In today's world, technology has become an integral part of the twenty-first century. The internet is one of these technologies, which is growing rapidly every year and plays an important role in individuals' lives. It has become a valuable and a convenient mechanism for supporting public transactions such as e-banking an de-commerce transactions. That has led the users to trust it is Convenient to provide their private information to the Internet. As a result, the security thieves that have started to target this a social engineering trick, which can be described as fraudsters that try to manipulate the user into giving them their personal information based on exploiting human vulnerabilities rather than software vulnerabilities.

Statistics have shown that the number of phishing attacks keeps increasing, which presents a security risk to the user information according to the Anti-Phishing Working Group(APWG) and recorded phishing attacks by Kaspersky Lab, which stated that it has increased by 47.48% from all of the phishing attacks that have been detected during2016.

Recently, there have been several studies that tried to solve the phishing problem. Some researchers used the URL and compared it with existing blacklists that contain lists of malicious websites, which they have been creating, and there are others that have used the URL in an opposite manner, namely comparing the URL with a whitelist of legitimate websites. The latter approach uses heuristics, which uses a signature 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 using Alexa is another way that has been implemented by researchers to detect phishing websites. Moreover, other researchers have used machine learning techniques.

Phishing is one in all the foremost problems of the knowledge security. It can occur in two ways, either by receiving suspicious emails that cause the fraudulent site or by users accessing links that go on to a phishing website. However, the 2 methods are common in one thing, which is that the attacker targets human vulnerabilities instead of software vulnerabilities. Phishing will be described as fraudsters that try and manipulate the user into giving them their personal information like username, password, and a master card number. These scams are resulting in economic and financial crises for users. within the early 90s, phishers created a false account with a fake identity and pretend master card on the America Online (AOL) company that provided an internet portal and was an online service provider. during this way, the phishers may be exploiting its services with none cost to them. Since then, within the mid 90s, AOL strengthened its system to stop phishers.

In this paper, the focus will be on the features combination that we get from Random Forest (RF) technique, as it has high accuracy, is relatively robust, and has a good performance.

II. II.LITERATURE SURVEY

[1] Title: Random Forest Explorations for URL Classification

Authors: Martyn Weedon, Dimitris Tsaptsinos, James Denholm- Price

Year of Publication: 2017

Abstract: Phishing is also a significant concern on the web today and lots of of users are falling victim due to criminal's

deceitful tactics. Blacklisting remains the foremost common defense users have against such phishing websites, but is failing to deal with the increasing number. In recent years, researchers have devised modern ways of detecting such websites using machine learning. One such method is to make machine learnt models of URL features to classify whether URLs are phishing. However, there are varying opinions on what the foremost effective approach is for features and algorithms. during this paper, the target is to judge the performance of the Random Forest algorithm employing a lexical only dataset. The performance is benchmarked against other machine learning algorithms and additionally against those reported within the literature. Initial results from experiments indicate that the Random Forest algorithm performs the foremost effective yielding an 86.9% accuracy.

[2] Title: A Hybrid Model to Detect Phishing-Sites using Supervised Learning Algorithms

Authors: M. Amaad Ul Haq Tahir, Sohail Asghar, Ayesha Zafar, Saira Gillani

Year of Publication: 2016

Abstract: Since last decades, online technologies have revolutionized the fashionable computing world. However, as a result, security threats are increasing rapidly. An enormous community is using the net services even from chatting to banking is completed via online transactions. Customers of web technologies face various security threats and phishing is one in every of the foremost important threat that has to be address. Therefore, the protection mechanism must be enhanced. The attacker uses phishing attack to induce victims credential information like checking account number, passwords or the other information by mimicking a web site of an enterprise, and also the victim is unaware of phishing website. In literature, several approaches are proposed for detection and filtering phishing attack.

However, researchers are still looking for such an answer that may provide better results to secure users from phishing attack. Phishing websites have certain characteristics and patterns and to spot those features can help us to detect phishing. to spot such features may be a classification task and may be solved using data processing techniques. during this paper, we are presenting a hybrid model for classification to beat phishing-sites problem. The experimental results showed that our proposed hybrid terms of high accuracy and fewer error rate.

[3] Title: Machine Learning Based Phishing websites Detection.

Authors: Huu Hieu Nguyen and Duc Thai Nguyen Year of Publication:2015

Abstract: Phishing could be a major problem that involves websites and fraudulent emails that aim to reveal users important information like financial data, emails, and other private information. Phishing activities are within the increasing trend, and lots of unsuspecting users have fall en victims of those websites and fraudulent emails. This paper has analysed the evaluation and style of the features wont to detect and reduce any false activity. the chosen features not only rely upon the characteristics of the URL (Uniform Resource Locator), but also on the web site content. The TF-IDF algorithm is employed to calculate the highest keywords of the website content that's wont to extract one among the important features. The technique was evaluated on the dataset of 4.420 legitimate URLs and 5.389 phishing URLs. By considering features and evaluating using 5 classification algorithms, the resulting classifiers obtain 98.8 % accuracy on detecting phishing website URLs.

III. SYSTEM ANALYSIS AND SYSTEM DESIGN

Existing System:

In existing system , There are three phishing techniques are use:-

- Blacklist
- heuristic
- content analysis

Blacklist:- The blacklist compares the URL with an existing database that contains an inventory of phishing website URLs.

Heuristics:- The heuristics approach uses the signature databases of any known attacks, to match it with the signature of a heuristic pattern.

Content analysis:- It is a content-based approach in detecting phishing websites, using well-known algorithms like term frequency/inverse document frequency (TF-IDF). It analyses the text-based content of a page itself to make a decision whether the web site is phishing or not.

Additionally, measuring website traffic using Alexa is another method that has been implemented by researchers to detect phishing websites.

Disadvantages of Existing System:

- The trade-off of using heuristics is failing to detect novel attacks, because it is simple to bypass the signatures through obfuscation.

  Also, updating the signature database is slow considering the expansion of novel attacks, especially zero-day attacks .

- Because of the rapid increase of phishing websites, the blacklist approach has become inefficient to decide whether each URL may be a phishing website or not, and this type of delay can result in zero-day attacks from new phishing sites .

PROPOSED SYSTEM:

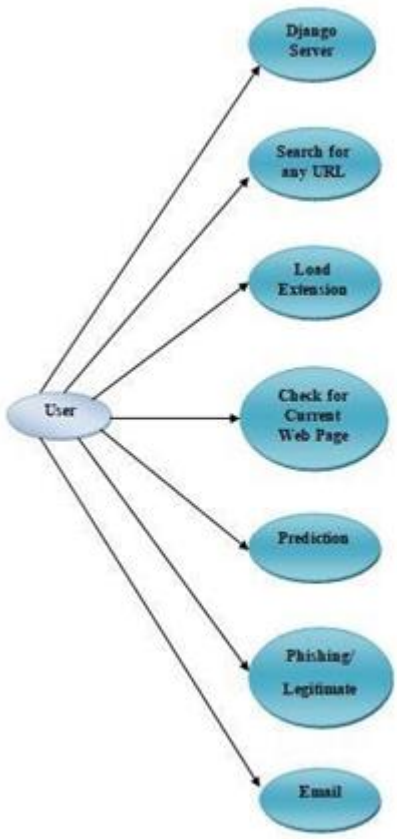
A new approach had proposed to detect phishing sites is by deriving different components from the URL and computing a metric for every component. Then, the page ranking are going to be combined with the achieved metrics to make a decision whether the websites are phishing websites. The results showed that the technique can detect over 97% of phishing websites. A system for prediction phishing URLs by generating rules of association rule mining. We used the Random Forest algorithm to choose known information from frequent item set properties that were extracted from the dataset and also used different algorithm that performs on hidden data to get the accuracy of association rules, which may be a predictive that engages the boldness and also the support techniques that are measured in its accuracy, unlike a priori, which only mark rules that have the boldness technique. As a result, they presented significant 16 features of the URL that distinguish if it's phishing or legitimate.

Advantages of Proposed System:

Performance and accuracy is more comparing to other similar application.
Reduce the Time Complexity

IV.IMPLIMENTATION

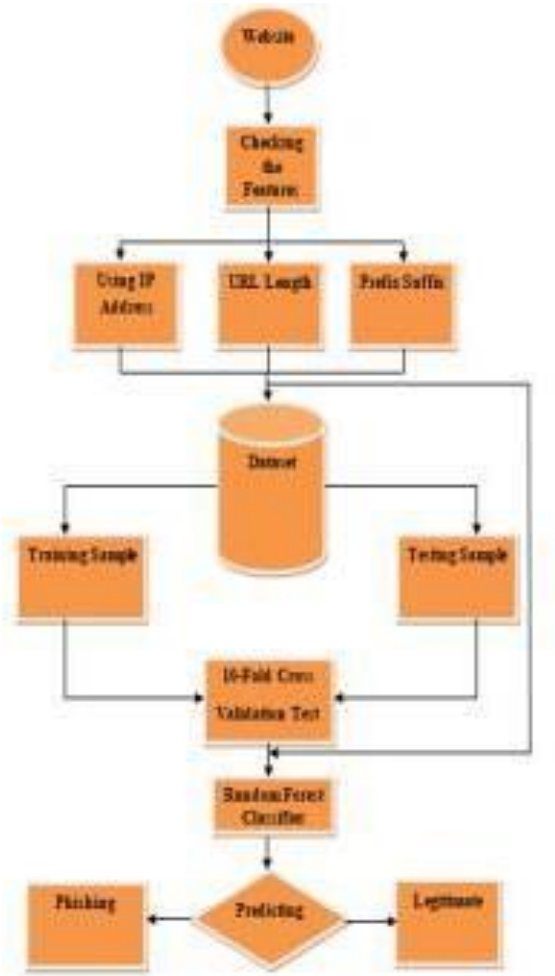
System Architecture:



A. Dataset We collect 16000 of phishing and legit URLs. The phishing websites encompass 12000 phishing URLs that has been collected from PhishTank. within the other hand, the legitimate websites encompass 4000 legitimate URLs that are collected by a daily use from 10 chosen users. However, the ultimate dataset after handling missing data and removing the duplicate is size of 6116.
B. Features extraction The phishing websites have certain characteristics and pat- terns that may be considered as features. during this subsection, we cover all phishing website features that are used within the previous researches as possible. Furthermore, while we are studying the phishing characteristics and patterns we notice some new characteristics that may be considered as features. the full number of phishing features is 36 where 3 of them are new features. We categorize them into three main categories.
   - Features will be extracted from URL.
   - Features will be extracted from page content.
   - Features will be extracted from page rank.

We use the quantity of input email and number of input password because the new features for phishing website, Since the target of the phishing website is to steal sensitive information like email and password. We consider the quantity of input that have the kind email or password as feature for phishing website. Another new feature is that the number of button, while we are studying phishing features we noticed that an oversized number of phishing website doesn't use the submit button instead they use an everyday button, so we consider it as feature for phishing website.

Data Flow Diagram:



V. METHODOLOGY

We study all features to point the strongest, weakest and to get rid of the irrelevant features; the study is predicated on examining all possible combination of 26 features. where k is that the number of the taken features that start from 1 to 26. and n is that the number of all features which is 26.

Since the quantity of all possible combination could be a huge number, the study are going to be summarized into taking the utmost and therefore the minimum result for every k combination. In the end, the upper accuracy with the tiniest number of features are going to be chosen for a much better combination. In Fig. 1, it summarizes the method of feature selection.

The main function of the system is to choose the state of the web site if it's a phishing or legitimate website. This function will be performed using the algorithm as shown in Fig. 2. This algorithm are going to be triggered whenever the user enters a brand new website, the role of the algorithm is to extract the features of the web site using URL and Document Object Model (DOM) object. The URL accustomed extract the URL's and page rank's features. While the DOM accustomed extract the content page's features which could be a connection between scriptsand website's page that have logical structure of documents and supply accessing and manipulation for programmer to the DOM file. Afterwards, the extracted features are going to be sent to the classifier to provide the target label that indicates the state of the web site then executes the acceptable action on it.

We build the classifier using RF technique as within the following steps:

1) Split data into training and test dataset, which we take 80% for training and 20% for testing.

2) Train and test all possible combination of 26 features dataset to induce the strongest features that arise the accuracy of detection.

3) After step two, we've numbers of features which works to the ultimate stage of coaching and testing.

4) Execute the ultimate classifier

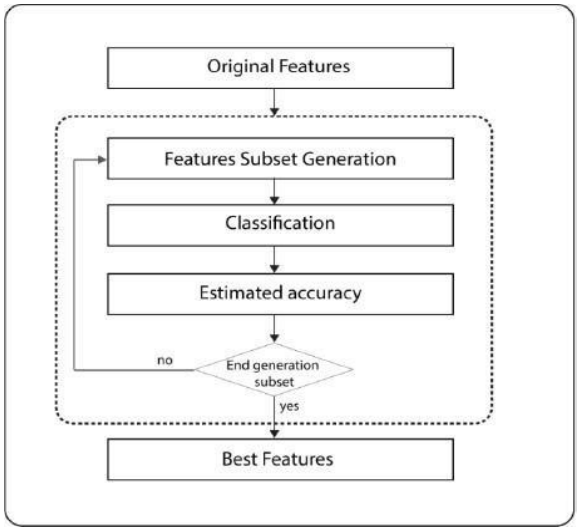


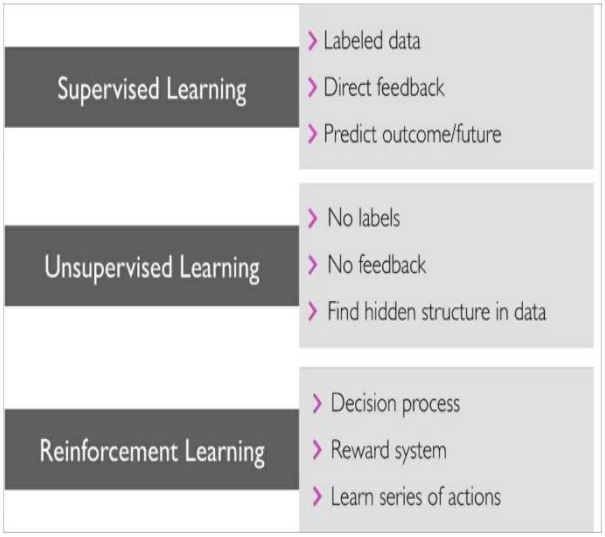
Fig 3 The process of feature selection

Phishing attacks being on a rise, using a direct search from the phishing website database is not enough. To provide protection against new attacks, machine learning provides the best alternative for the same. We used the UCI Dataset of Phishing Website to train the classifier. Later, whenever a user enters the URL, the features are extracted and the URL is tested on the trained classifier to obtain the result.

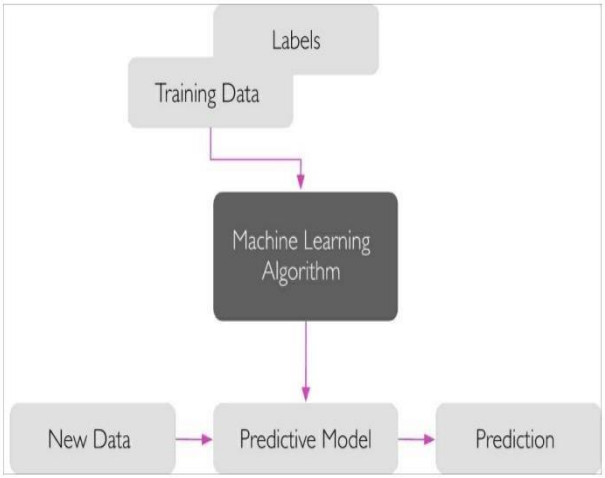
Machine learning using python: In this age of modern technology, there is one resource that we have in abundance: a large amount of structured and unstructured data. In the second half of the twentieth century, machine learning evolved as a subfield of Artificial Intelligence (AI) that involved selflearning algorithms that derived knowledge from data in order to make predictions. Instead of requiring humans to manually derive rules and build models from analyzing large amounts of data, machine learning offers a more efficient alternative for capturing the knowledge in data to gradually improve the performance of predictive models and make data-

driven decisions. Not only is machine learning becoming increasingly important in computer science research, but it also plays an ever greater role in our everyday lives. Thanks to machine learning, we enjoy robust email spam filters, convenient text and voice recognition software, reliable web search engines, challenging chess-playing programs, and, hopefully soon, safe and efficient self-driving cars.

we will take a look at the three types of machine learning: supervised learning, unsupervised learning, and reinforcement learning. We will learn about the fundamental differences between the three different learning types and, using conceptual examples, we will develop an intuition for the practical problem domains where these can be applied:



Making predictions about the future with supervised learning:

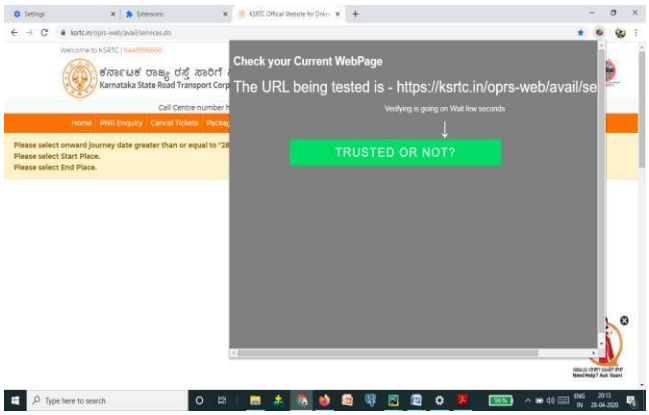


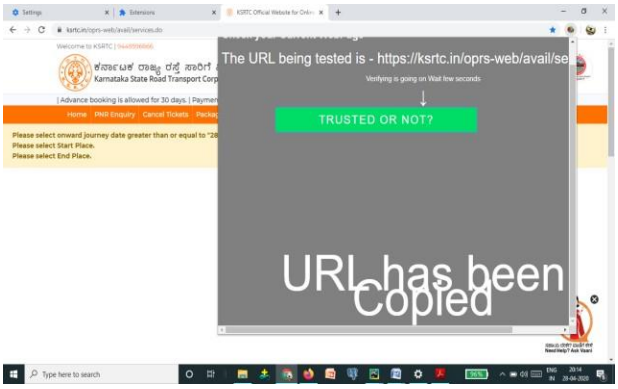
The main goal in supervised learning is to find out a model from labeled training data that enables us to form predictions about unseen or future data. Here, the term supervised refers to a group of samples where the specified output signals (labels) are already known.

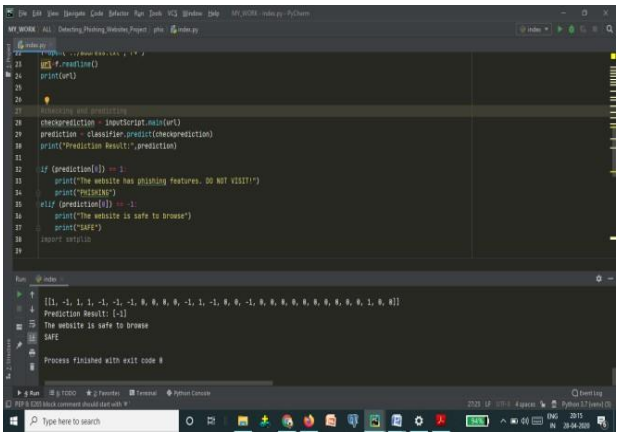
Considering the instance of email spam filtering, we will train a model employing a supervised machine learning algorithm on a corpus of labeled emails, emails that are correctly marked as spam or not-spam, to predict whether a replacement email belongs to either of the 2 categories. A supervised learning task with discrete class labels, like within the previous email spam filtering example, is additionally called a classification task. Another subcategory of supervised learning is regression, where the end result signal may be a continuous value.

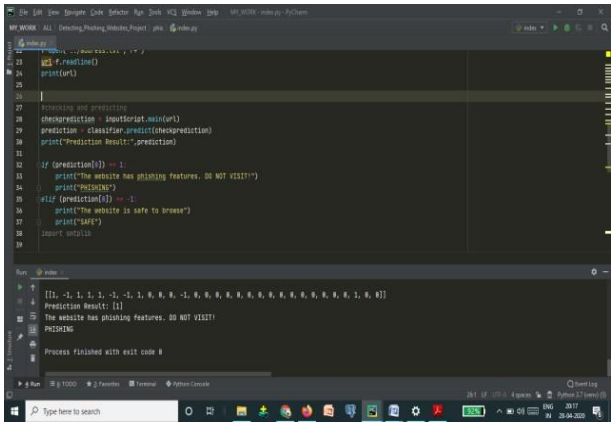
Python is one amongst the foremost popular programming languages for data science and thus enjoys an oversized number of useful add-on libraries developed by its great developer and and open-source community. Although the performance of interpreted languages, like Python, for computation-intensive tasks is inferior to lower-level programming languages, extension libraries like NumPy and SciPy are developed that depend upon lower-layer Fortran and C implementations for fast and vectorized operations on multidimensional arrays. For machine learning programming tasks, we will mostly discuss with the scikitlearn library, which is currently one amongst the foremost popular and accessible open source machine learning libraries.

Python is one amongst the foremost popular programming languages for data science and thus enjoys an oversized number of useful add-on libraries developed by its great developer and and open-source community. Although the performance of interpreted languages, like Python, for computation-intensive tasks is inferior to lower-level programming languages, extension libraries like NumPy and SciPy are developed that depend upon lower-layer Fortran and C implementations for fast and vectorized operations on multidimensional arrays. For machine learning programming tasks, we will mostly discuss with the scikitlearn library, which is currently one amongst the foremost popular and accessible open source machine learning libraries.









VI. CONCLUSIONS

In this paper, we defined features of phishing attack and we proposed a classification model in order to classification of the phishing attacks. This method consists of feature. extraction from websites and classification section. In the feature extraction, we have clearly defined rules of phishing feature extraction and these rules have been used for obtaining features.

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/press-releases/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.