**DETECTION OF PHISHING WEBSITES FROM URL’S**

**USING IBM WATSON STUDIO**

**TEAM MEMBERS:**

* Sumasri Dandi
* Shiva Sripelly
* Srisairam Govindula
* Keshav Ellaboina

**INTRODUCTION**

1. **1 OVERVIEW:**

Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Logistic Regression, Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

**1.2 PURPOSE:**

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. In United States businesses, there is a loss of US$2billion per year because their clients become victim to phishing [1]. In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $5 billion [2]. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques. The general method to detect phishing websites by updating blacklisted URLs, Internet Protocol (IP) to the antivirus database which is also known as “blacklist" method. To evade blacklists attackers uses creative techniques to fool users by modifying the URL to appear legitimate via obfuscation and many other simple techniques including: fast-flux, in which proxies are automatically generated to host the web-page; algorithmic generation of new URLs; etc. Heuristic based detection which includes characteristics that are found to exist in phishing attacks in reality and can detect zero-hour phishing attack, but the characteristics are not guaranteed to always exist in such attacks and false positive rate in detection is very high [3]. To overcome the drawbacks of blacklist and heuristics based method, many security researchers now focused on machine learning techniques. Machine learning technology consists of a many algorithms which requires past data to make a decision or prediction on future data. Using this technique , algorithm will analyze various black listed and legitimateURLs and their features to accurately detect the phishing websites.

**2 . LITERATURE SURVEY**

**2.1 EXISTING PROBLEM**

The current problem is Phishing. These social engineering attacks are designed to **fool you** into causing a data breach. Phishing attackers pose as people or organizations you trust to easily deceive you. Criminals of this nature try to coax you into handing over access to sensitive data or provide the data itself.

**2.2 PROPOSED SOLUTION**

We have implemented python program to extract features from URL. Below are the features that we have extracted for detection of phishing URLs.

* **Presence of IP address in URL**: If IP address present in URL then the feature is set to 1 else set to 0. Most of the benign sites do not use IP address as an URL to download a webpage. Use of IP address in URL indicates that attacker is trying to steal sensitive information.
* **Presence of @ symbol in URL:** If @ symbol present in URL then the feature is set to 1 else set to 0. Phishers add special symbol @ in the URL leads the browser to ignore everything preceding the “@” symbol and the real address often follows the “@” symbol.
* **Number of dots in Hostname:** Phishing URLs have many dots in URL. For example http://shop.fun.amazon.phishing.com, in this URL phishing.com is an actual domain name, whereas use of “amazon” word is to trick users to click on it. Average number of dots in benign URLs is 3. If the number of dots in URLs is more than 3 then the feature is set to 1 else to 0.
* **Prefix or Suffix separated by (-) to domain:** If domain name separated by dash (-) symbol then feature is set to 1 else to 0. The dash symbol is rarely used in legitimate URLs. Phishers add dash symbol (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example Actual site is http://www.onlineamazon.com but phisher can create another fake website like http://www.online-amazon.com to confuse the innocent users.
* **URL redirection:** If “//” present in URL path then feature is set to 1 else to 0. The existence of “//” within the URL path means that the user will be redirected to another website.
* **HTTPS token in URL:** If HTTPS token present in URL then the feature is set to 1 else to 0. Phishers may add the “HTTPS” token to the domain part of a URL in order to trick users. For example, http://https-wwwpaypal-it-mpp-home.soft-hair.com.
* **Length of Host name:** Average length of the benign URLs is found to be a 25, If URL’s length is greater than 25 then the feature is set to 1 else to 0.
* **Presence of sensitive words in URL:** Phishing sites use sensitive words in its URL so that users feel that they are dealing with a legitimate webpage. Below are the words that found in many phishing URLs :- 'confirm', 'account', 'banking', 'secure', 'ebyisapi', 'webscr', 'signin', 'mail', 'install', 'toolbar', 'backup', 'paypal', 'password', 'username', etc;
* **Number of slash in URL:** The number of slashes in benign URLs is found to be a 5; if number of slashes in URL is greater than 5 then the feature is set to 1 else to 0.
* **Presence of Unicode in URL:** Phishers can make a use of Unicode characters in URL to trick users to click on it. For example the domain “xn--80ak6aa92e.com” is equivalent to "аррӏе.com".

Visible URL to user is "аррӏе.com" but after clicking on this URL, user will visit to “xn--80ak6aa92e.com” which is a phishing site.

* **Website Rank:** We extracted the rank of websites and compare it with the first One hundred thousand websites of Alexa database. If rank of the website is greater than 10,0000 then feature is set to 1 else to 0.

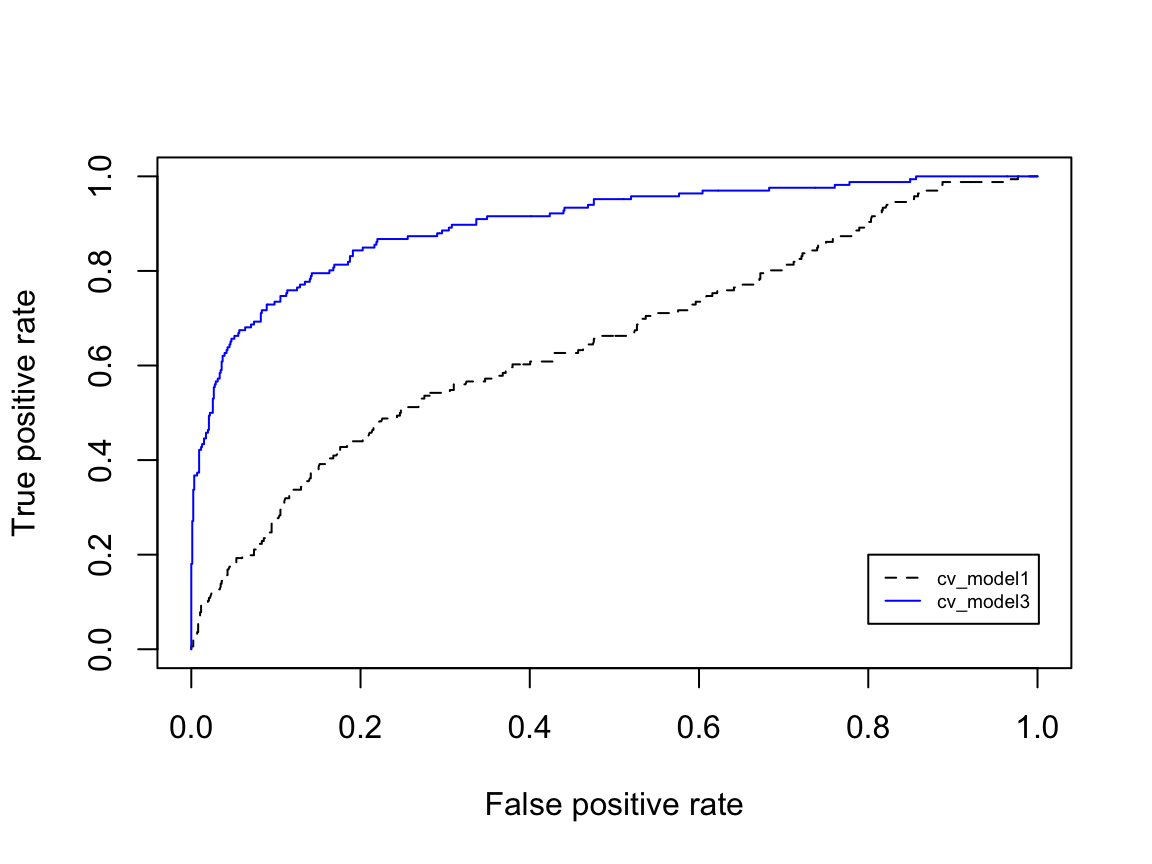
1. **THEORITICAL ANALYSIS**

3.1 **BLOCK DIAGRAM**

One machine learning model, Logistic Regression has been selected to detect phishing websites.

**Logistic Regression:** Logistic Regression is a Machine Learning algorithm which is used for the **classification problems**, it is a predictive analysis algorithm and based on the concept of probability. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Logistic regression is a simple yet very effective classification algorithm so it is commonly used for many  **binary classification tasks**. Customer churn, spam email, website or ad click predictions are some examples of the areas where logistic regression offers a powerful solution.

Logistic regression uses an **equation as the representation**, very much like linear regression. Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y).



**FIG:** Logistic regression model comparision.

* 1. **HARDWARE/ SOFTWARE DESIGNING**

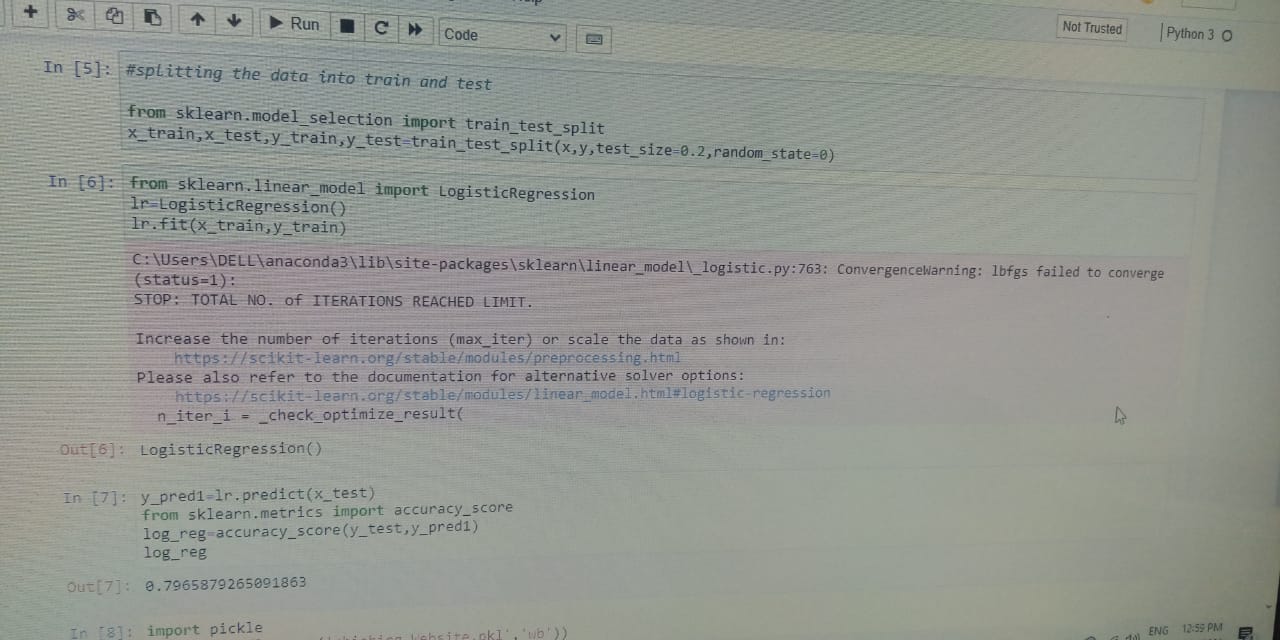
**Hardware Requirements:**

* Processor Minimum: Minimum 1GHz; Recommended 2GHz or more
* Ethernet Connection(LAN) or a Wireless adapter (WiFi)
* Hard Drive: Minimum 32GB ; Recommended 64 GB or more
* Memory (RAM): Minimum 1GB; Recommended 4GB or above

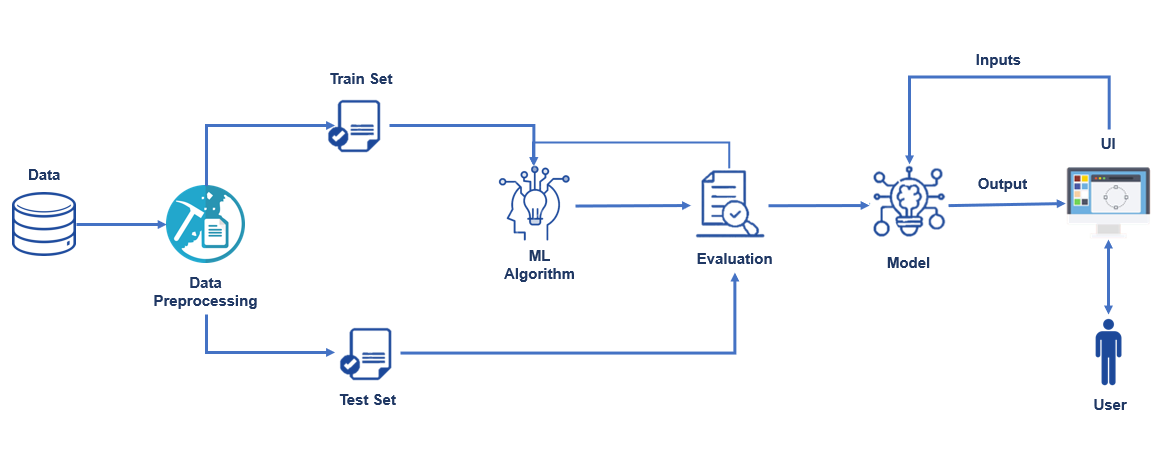
**Software Requirements:**

* Python(3.7 or older)
* Anaconda Prompt
* Watson Studio Service
* Spyder 3.8
* Flask
* Jupiter Notebook

1. **EXPERIMENTAL INVESTIGATION**

****

1. **FLOWCHART:**



1. **RESULT:**

Scikit-learn tool has been used to import Machine learning algorithms. Dataset is divided into training set and testing set . The classifier is trained using training set and testing set is used to evaluate performance of classifier. Performance of classifier has been evaluated by calculating classifier's accuracy score, false negative rate and false positive rate.

Result also shows that detection accuracy of phishing websites increases as more dataset used as training dataset. All classifiers perform well when 90% of data used as training dataset.

Fig. 1 show the detection accuracy of the classifier when 50%, 70% and 90% of data used as training dataset and graph clearly shows that detection accuracy increases when 90% of data used as training dataset.

1. **ADVANTAGES:**

* Eliminate the cyber threat risk level.
* Increase user alertness to phishing risk.
* Instill a cyber security culture and create cyber security heroes.
* Change behavior to eliminate the automatic trust response.

DISADVANTAGES:

* Employees and customers inability to detect phishing emails and messages
* Insufficient communication between management organizations and employees or customers

1. **APPLICATIONS:**

* This system will be useful for many E-Commerce enterprises.
* This system will be useful for many users who purchase products online.

9. **CONCLUSION**

This paper aims to enhance detection method to detect phishing websites using machine learning technology. We achieved 97.14% detection accuracy using Logistic regression algorithm with lowest false positive rate. Also result shows that classifiers give better performance when we used more data as training data. In future hybrid technology will be implemented to detect phishing websites more accurately, for which logistic regression algorithm of machine learning technology and blacklist method will be used.

1. **FUTURE SCOPE:**

* Phishing attacks in the future **could take multiple forms and could evolve beyond recognition**. For right now, your enterprise needs phishing protections such as email security to prevent the majority of phishing attacks from ever reaching your employees in the first place.

11. **BIBLOGRAPHY**

* Gunter Ollmann, “The Phishing Guide Understanding & Preventing Phishing Attacks”, IBM Internet Security Systems, 2007.
* <https://github.com>
* <https://smartinternz.com>
* www.google.com