Problem Statement – Web Phishing Detetion

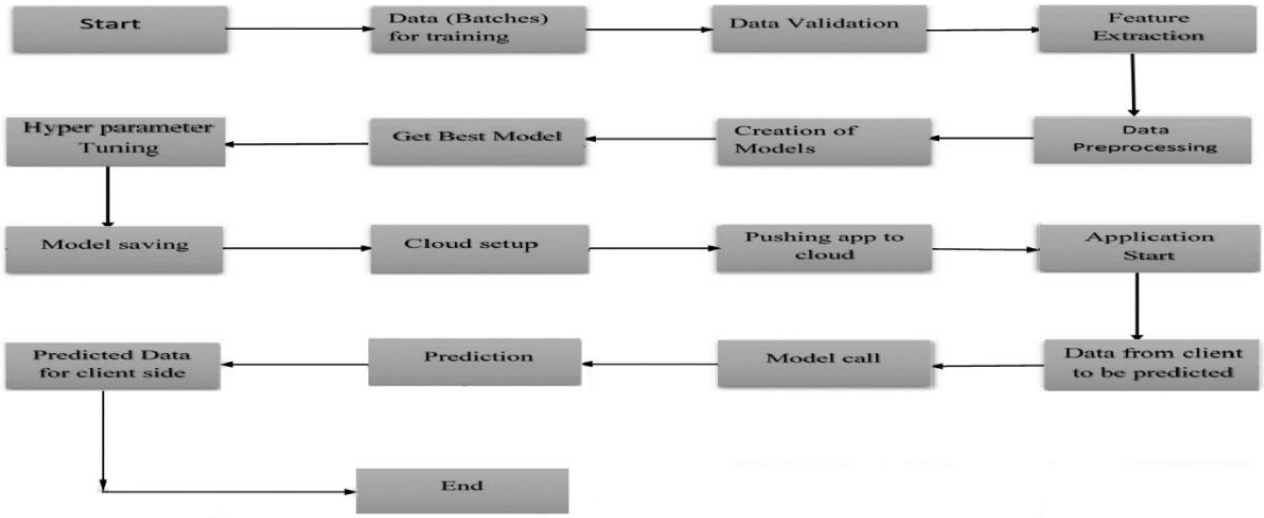
Internet has dominated the world by dragging half of the world's population exponentially into the cyber world. With the booming of internet transactions, cybercrimes rapidly increased and with anonymity presented by the internet, Hackers attempt to trap the end-users through various forms such as phishing, SQL injection, malware, man-in-the-middle, domain name system tunnelling, ransomware, web trojan, and so on. Among all these attacks, phishing reports to be the most deceiving attack. Our main aim of this paper is classification of a phishing website with the aid of various machine learning techniques to achieve maximum accuracy and concise model.

Motivation

Detection and prevention of phishing websites endure measure continuously a major space for analysis. There are different types of phishing techniques that offer torrential and essential ways that offer attackers to penetrate the data of people and organizations. Uniform resource locator URLs sometimes are also referred to as "Weblinks" play a vital role in a phishing attack. Uniform resource locator has a vulnerability of redirecting the pages i.e., through the hyperlink; which could redirect to the legitimate website or the phishing site. Different techniques in making phishing sites are emerging day by day. This actually motivated several researchers to put up their concentrate on finding the phishing sites.

Data flow

The technique comprises of host based, page based and lexical feature extraction of collected websites. The primary step is the collection of phishing and benign websites. In the host-based approach, admiration based and lexical based attributes extractions are performed to form a database of attribute value. This database consists of knowledge mined that uses different machine learning techniques. On evaluating the algorithms, a selective classifier is opted and is implemented in Python.



URL collection

We collected URLs of benign websites from www.alexa.com ,www.dmoz.org and personal web browser history. The phishing URLs were collected from www.phishtak.com [8]. The data set consists of 17000 phishing URLs and 20000 benign URLs. We obtained PageRank of 240 benign websites and 240 phishing websites by checking PageRank individually at PR Checker. We collected WHOIS information of 240 benign websites and 240 phishing websites.

Host-based analysis

Host-based features explain “where” phishing sites are hosted, “who” they are managed by, and “how” they are administered. We use these features because phishing Web sites may be hosted in less reputable hosting centres, on machines that are not usual Web hosts, or through not so reputable registrars.

Below are the characteristics of the host-based that are notified.

i. WHOIS properties: WHOIS properties give information regarding the registrations, updates and expiry, differentiating the admin and the user. Phishing URLs are taken down repeatedly, the date of registration will be recent compared to legitimate sites. Majority of phishing URLs contain IP address in their hostname.

ii. Geographic properties: Geographic properties provides the information regarding the continent/state/country to which the corresponding IP address belongs to. Analyse attributes using machine learning techniques. Selection of a

Classifier Implement the classifier stores phishing & Benign websites host-based and page-based attribute Lexical feature extraction.

Machine learning algorithms

The evaluation of the various classifying algorithm is done by using the workbench for data mining, Waikato Environment for Knowledge Analysis (WEKA) [13-16]. Four types of input data files i.e., Attribute Relation File Format (.arff), Comma Separated Values (.csv). In our experiment .csv file format was used. The input file to the Waikato Environment for Knowledge Analysis was obtained by program by appending 'YES' in place of decision vector '1' (phish) and 'NO' in place of decision vector '0' (benign) of the dataset generated from input URL list. The dataset was made split into 70% for training and remaining 30% for testing purpose.

The five machine learning algorithms considered for processing the feature set are:

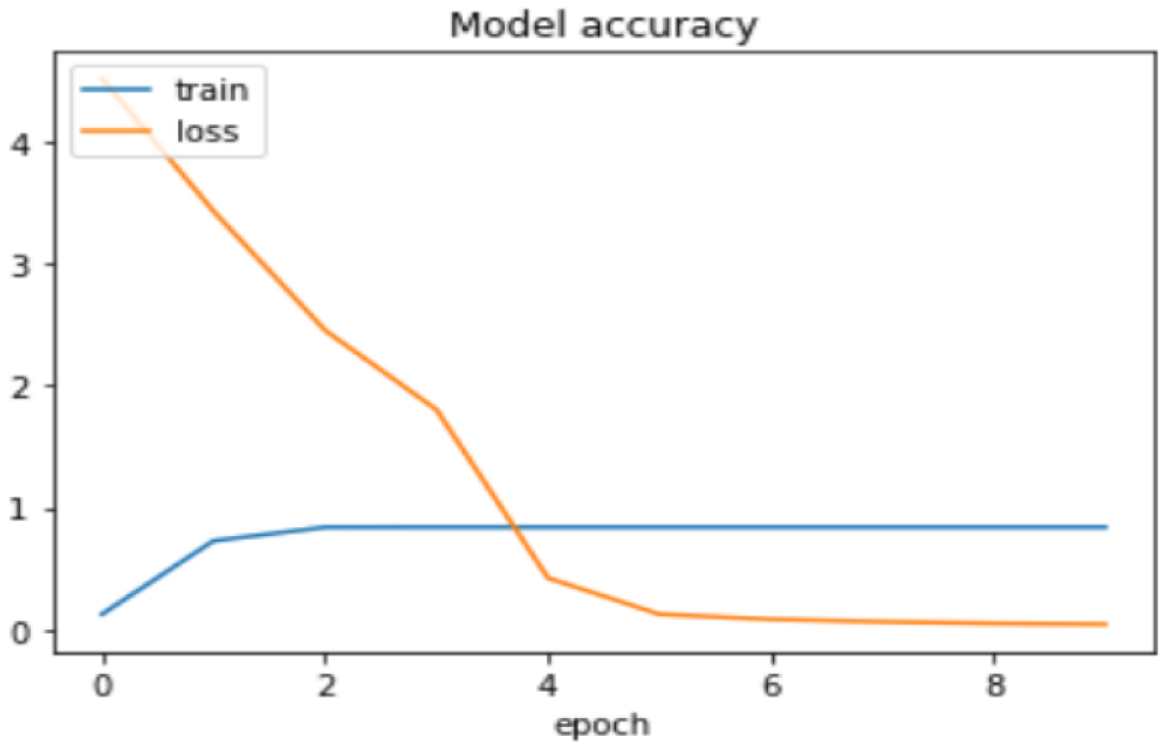
1) Logistic regression: It is a statistical model that uses a logistic function to build a dependent variable, which can also have many more complex extensions.

2) SVM: The Support Vector Machine performs a classification task by finding the 'hyper plane' which maximizes the margin between two groups of classes. The vectors that signify the hyper plane are known as the support vectors.

3) XGBoost: Boosting is a machine learning technique used in regression, classification and other tasks, that predicts a model in the form of an ensemble prediction models, favourably decision trees.

4) MLP: A Multilayer Perceptron (MLP) is a class neural network typically Artificial Neural Networks (ANN). Intuitively known as Perceptron of multiple layers.

5) AutoEncoders: An autoencoder is a type of neural network typically Artificial Neural Networks (ANN) that is used to learn the patterns of the unlabelled data (unsupervised learning). The Figure 5 shows the loss function of the autoencoders model. Binary crossentropy is the loss function that is defined on the training data and in Figure 5, the blue line represents the training loss and the orange line represents the accuracy of AutoEncoders.



Results

- The key notable points of our initial work embed:
- Phishing sites and their domains reveal the features that are different from other sites and domains. (For example, Google; www.google.com and some random phishing website be like; www.googlee.com).
- Phishing Uniform Resource Locators and ‘domain names’ typically have a different length when compared to other websites and domain names.

Conclusion

Phishing is a major problem, which uses both social engineering and technical deception to get users’ important information such as financial data, emails, and other private information. Phishing exploits human vulnerabilities; therefore, most protection protocols cannot prevent the whole phishing attacks. Many of them use the blacklist/whitelist approach, however, this cannot detect zero-hour phishing attacks, and they are not able to detect new types of phishing attacks.