IJARIIT International Journal Of Advance Research, Ideas And Innovations In Technology

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 3 - V7I3-2143)
Available online at: https://www.ijariit.com

Detection and Classification of Malicious Websites

Karan Sawhney
badman8991@gmail.com
RV College of Engineering, Bengaluru, Karnataka

Priya D.
priyad@rvce.edu.in
RV College of Engineering, Bengaluru, Karnataka

ABSTRACT

Phishing is an attractive methodology that is used by the attackers to steal personal data from people. Phishing websites are fraudulent web pages used for cyber thefts. They use domain names, similar to that of an original website, send mail and direct messages to lure someone in sharing personal information. Although many new anti-phishing methods are out ,still these attackers innovate new methods to target these. Therefore there is an alarming need to develop new ways to protect people from phishing. The Project uses different models of comparison in classification of phishing websites for detection and prediction using different machine learning models. The main goal of the project is to make Machine Learning Algorithm for data input and prediction and classify URL's that are malicious . Using python packages(Fastai, Scikit,pandas, Seaborn etc.) the dataset is converted into Distribution Plots and Confusion Matrix.

Keywords: *Seaborn, Classification and Regression Trees (CART)*

1. INTRODUCTION

In the last few years, machine learning methods have been applied for a wide range of applications. The use of machine learning in information theory, it is and always has been. Machine learning algorithms have a lot in common with the bank. Machine learning can be used in the training and testing of large amounts of data, and the major search engines in the room. It is important to cover the ML techniques in the algorithms.

Machine learning methods have been used in different research areas with high potential. These applications are in the areas of cryptography and cryptanalysis, steganography, and data security-related applications. Machine learning makes use of analytical models to be used, and, to use a large amount of data have to be entered as an introduction. Machine learning techniques can be used to show the relationship between the input and output of data that can be used by the cryptosystems.

2. STUDY ON RELATED WORK

Jakobbson and Meyers 2007. The paper had a research mainly to analyze and detect the different features that were involved with the online scam and URL intrusions. [8]

D.Sahoo, C.Liu, and S.C.H. Hoi [2], In order to provide an in-depth study of the structural concept, the malicious URL detection methods, with the help of machine learning. We present you with the official formulation of the URLS, such as machine learning tasks, as well as the classification and the analysis of the contributions of literary studies, which have an influence on the various aspects of a particular issue (the concepts, features, algorithm development, etc.). In addition, this article will give you a quick and comprehensive overview for a variety of recipients, not only for machine learning researchers and engineers in academia, but also for cyber security professionals in order to help them to understand the current situation and to facilitate their private study and for practical applications.

3. OVERVIEW OF THE TECHNIQUES

Logistic Regression

A logistic regression model, a learning algorithm that takes into account the probability of any particular variable. This variable is the only one way to tell us that there are only two possible classes. Simply put, the dependent variable is binary data, which is either 1(success) or 0(failure).

K-Nearest Neighbor

The k-Nearest Neighbor Algorithm (KNN) is a simple supervised machine learning algorithm which can be used to solve classification and regression problems. It is easy to implement and understand, but it has one major drawback-it slows down significantly the amount of data transmitted increases.

Classification and Regression Tree

Classification and Regression Trees (CART), is a prediction algorithm that is used in machine learning model. This shows a prediction of the target group of the dependent variable on the basis of different values. CART is a decision tree, where each module has been divided into a variable and each of the nodes in the end, it is a prediction of the variable.

Support Vector Machine

The Support Vector Machine (SVM) is supervised by a machine learning algorithm that can be used for both classification and regression tasks. However, it is mostly used in classification tasks. SVM algorithm is to construct every element of data as a point in n-dimensional space (where n is the number of objects that you already have), and the value of each and every object is the value of a particular town or city.

Decision Trees

Decision tree is a decision-making tool that uses a tree-like graph or model to represent decisions and their potential results, such as chance event outcomes, resource costs, and more. It's one method of displaying an algorithm with just conditional power.

3. METHODOLOGY

This chapter focuses on the methods and techniques used for the detection of phishing, and identification of a set of data with the help of the selection.A web site is a social engineering technique, which presents itself as a reliable, uniform resource locators (Urls) and for the website. The project goal is to collect data and to delete it from the Url. The project will make use of the Feature Selection in the machine learning algorithms, which explains in detail. Have a look at some of the parameters that define the network, as will be shown below.

3.1. Dataset Collection

A set of phishing sites, which can easily be reached by means of the open-source services. The service provides a set of phishing sites in a variety of formats, such as csv, json, etc. Legitimate web sites, we make use of the open-source code that contains a collection of lightweight, spam, phishing attacks, malicious and bad types of web sites. It's the number of well-known sites in this collection of 35,300. We consider a margin value of 5000 phishing websites and 5000 benign websites.

3.2 Features Extracted

The features extracted from the domain bar are Domain of URL, IP Address in URL, "@" symbol in URL, Length of URL, Depth of URL, Redirection"//" in URL, "http/https" in Domain Names, Using Short URL, Prefix or Suffix '-'in domain etc.

3.3 Final Selection

In this section, click to select the features-accept frames from both legitimate and phishing websites. The two frames can be combined into a single data frame and the data can be exported as a histogram for a data visualization.

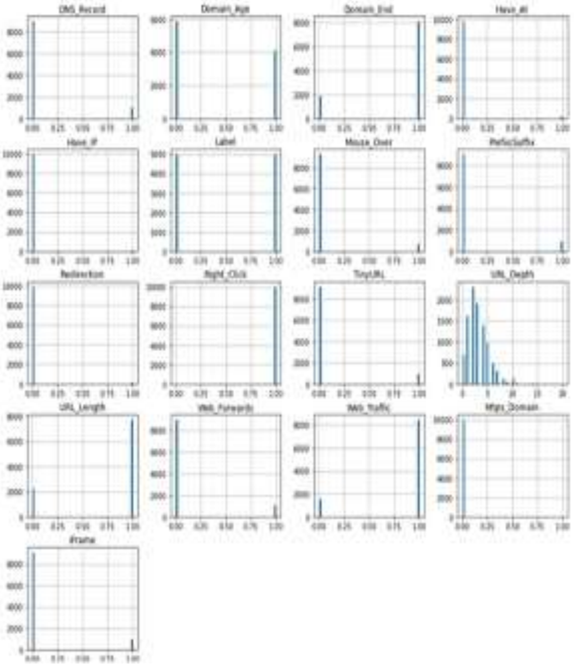

Fig. 1: Feature and Selection of Attributes.

3.1.4 Dataset Visualization

Heatmap to visualize the 18extracted for 10k URL are correlated on a heatmap.

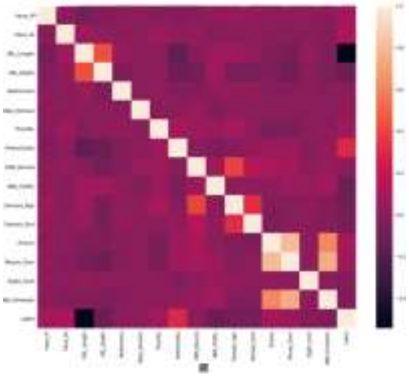

Fig. 2: Heatmap

4. COMPREHENSIVE ANALYSIS

The algorithms can also be used to improve the learning rate of the model. The learning rate is enhanced by the variation in the different stages of TDLHBA of the parameters, and the entity's implementation, and a single cycle of policy-making. Increase the learning rate to be used in order to consider the best model. The ML model is used to predict and classify malicious web sites. Untitled It is clear from our data set, we are in need of a supervised machine learning task. The two main types of supervised machine learning tasks of classification and regression. The data set belong to the classification task, since the input to the site, it is classified as a phishing scam (1)& benign (0). The supervised machine learning tasks (classification) mention in the subject matter being studied, and the accuracy is evaluated.

5. CONCLUSION

The project has been able to adapt to a malicious site from randomly selected sites, with the help of machine learning algorithms and neural networks. The symptoms can be eliminated by the selection of the attributes on the model. The machine learning algorithm under the random forest classifier, we can improve the learning rate of the model. It is a model that will allow us to increase the detection accuracy. In the prediction process, you will be able to correctly predict all of the malicious, phishing sites, and more. It is better to make use of the random forest classifiers, since our dataset contains categorical and quantitative value, a random forest, does a better job. It's based on, the trees, the scale of the object. One-sided changes are easily caught by the trees. It is used in an arbitrary sub-range, and prevents the excess seen. It supports missing values, with a high score. Large volumes of three-dimensional space have been processed.

Table1. Model Evaluation

Model	Training Accuracy	Testing Accuracy	Loss
Random Forest	99 %	98.6566 %	0.002187
Decision Tree	98 %	98.1298 %	0.004438
Classification & Regression Tree	98 %	98.1100 %	0.004848
K- Nearest Neighbours	97 %	97.0234 %	0.003176
Support Vector Machine	95 %	95.4758 %	0.005046
Logistic Regression	95 %	95.4363 %	0.004674

6. REFERENCES

[1] A. S. Manjeri, R. Kaushik, M. Ajay, and P. C. Nair, “A machine learning approach for detecting malicious websites using URL features,” in *2019 3rd International con- ference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, Jun. 2019. doi: 10.1109/iceca.2019.8821879.

[2] D.Sahoo, C.Liu, and S.C.H. Hoi, “Malicious url detection using machine learning: A survey,” Jan. 2017. arXiv: 1701.07179 [cs.LG].

[3] N. Singh, N. S. Chaudhari, and N. Singh, “Online URL classification for large-scale streaming environments,” *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 31–36, Mar. 2017. doi: 10.1109/mis.2017.39.

[4] A. K. Singh and N. Goyal, “A comparison of machine learning attributes for detect- ing malicious websites,” in *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*, IEEE, Jan. 2019. doi: 10.1109/comsnets. 2019.8711133.

[5] V.M.Patro and M.R.Patra, “Augmenting weighted average with confusion matrix to enhance classification accuracy,” *Transactions on Machine Learning and Artificial Intelligence*, vol. 2, no. 4, Aug. 2014. doi: 10.14738/tmlai.24.328.

[6] H. Kumar, P. Gupta, and R. P. Mahapatra, “Protocol based ensemble classifier for malicious URL detection,” in *2018 3rd International Conference on Contemporary Computing and Informatics (IC3I)*, IEEE, Oct. 2018. doi: 10.1109/ic3i44769. 2018.9007255.

[7] J. Yoon, W. R. Zame, and M. van der Schaar, “ToPs: Ensemble learning with trees of predictors,” *IEEE Transactions on Signal Processing*, vol. 66, no. 8, pp. 2141–2152, Apr. 2018. doi: 10.1109/tsp.2018.2807402.