Data Analytics and Artificial Intelligence
Vol: 2(5), 2022
REST Publisher; ISBN: 978-81-948459-4-2
Website: http://restpublisher.com/book-series/data-analytics-and-artificial-intelligence/

Web Phishing Detection Using Machine Learning

[*]Yoga Priyan M, Shree Bhargav RK
Adhiyamaan College of Engineering, Tamilnadu, India.
[*]Corresponding author Email: yogaprian1@gmail.com

Abstract. There are e-banking websites that ask users to provide sensitive data such as username, password & Credit card details, etc., often for malicious reasons. It will lead to information disclosure and property damage. Large organizations may get trapped in different kinds of scams. They may also lose its reputation of the organization. To detect and predict e-banking phishing websites, we proposed an intelligent, flexible and Effective system that's based on using classification algorithms. Decision tree classifiers and various algorithms will be used for making accurate predictions on phishing websites. A web service is one of the most important Internet communications software services. Using fraudulent methods to get personal information is becoming increasingly wide spread these days. However, it makes our lives easier, it leads to numerous security vulnerabilities to the Internet's private structure. Web phishing is just one of the many security risks that web services face. Phishing assaults are usually detected by experienced users however, security is a primary concern for system users who are unaware of such situations. Phishing is the act of portraying malicious web runners as genuine web runners to obtain sensitive information from the end-user. Phishing is currently regarded as one of the most dangerous threats to web security. Vicious Websites significantly encourage Internet criminal activity and inhibit the growth of Web services. As a result, there has been a tremendous push to build a comprehensive solution to prevent users from accessing such websites. We suggest a literacy-based strategy to categorize Web sites into three categories: begin, spam, and malicious. Our technology merely examines the Uniform Resource Locator (URL) itself, not the content of Web pages. As a result, it removes run-time stillness and the risk of drug users being exposed to cyber surfer-based vulnerabilities. When compared to a blacklisting service, our approach performs better on generality and content since it uses learning techniques. Keywords: - Phishing, classification algorithm, Threats to web security.

1. Introduction

There are a number of users who purchase products online and make payments through e-banking. There are e-banking websites that ask users to provide sensitive data such as username, password & credit card details, etc., often for malicious reasons. This type of e-banking website is known as a phishing website. Web service is one of the key communications software services for the Internet. Web phishing is one of many security threats toweb services on the Internet. To preserve confidentiality, to protect the user from phishing websites, to developa user-friendly environment, to prevent or mitigate harm or destruction of computer networks, applications, devices.

2. Literature Review

In this section, few of the research works that deploy the above-mentioned algorithms are reviewed and their results are summarized.

- Rishikesh Mahajan and Irfan Siddavatam chose three algorithms for classification–Decision Tree, Random Forest and Support Vector Machine. Their dataset contained 17,058 benign URLs and 19,653phishing URLs collected from Alexa website and Phish Tank respectively, with 16 features each. The dataset was divided into training and testing set in the ratios 50:50, 70:30 and 90:10 respectively. The accuracy score, false negative rate and false positive rate were considered as performance evaluation metrics. They achieved 97.14% accuracy for Random Forest algorithm with the lowest false negative rate. The paper concluded that accuracy increases when more data is used for training.
- Jitendra Kumar et al. in trained different classifiers like Logistic Regression, Naive Bayes Classifier, RandomForest, Decision Tree and K-Nearest Neighbor based on the features extracted from the lexical structure of the URL. They created the dataset of URLs in such a way that it solved the issues of data imbalance, biased training, variance and overfitting. The dataset contained an equal number of labeled phishing and legitimate URLs, and was further split in the ratio 7:3 for training and testing. All the classifiers had almost the same AUC (area under ROC curve), but the Naive Bayes Classifier turned out to be more suitable as it had the highest AUC value. Naive Bayes achieved the highest accuracy of 98% with a precision=1, recall=0.95 and F1-score=0.97.
- Mehmet Korkmaz et al. proposed in a machine-learning based phishing detection system by using 8 differentalgorithms on three different datasets. The algorithms used were Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), XGBoost,Random Forest (RF) and Artificial Neural Network (ANN). It was observed that the models using LR, SVMand NB have low accuracy rate. In terms of training time, NB, DT, LR and ANN algorithms gave better results. They concluded that RF or ANN Algorithm may be used

because of less training time along with a high accuracy rate. So these algorithms are preferred for machine learning based phishing detection.

3. Proposed System

We collect the dataset required to train our model from Kaggle. We split the dataset into 80% for trainingand20% for testing and evaluating our model. Next we pre-process the dataset by imputing missing values andperforming outlier treatment. We pass this dataset and train our model using SVM, Decision Tree, Random Forest, XG Boost Algorithm-CatBoost Classifier. Finally we perform Stacking of above algorithms to predict it's a Phishing site or not.

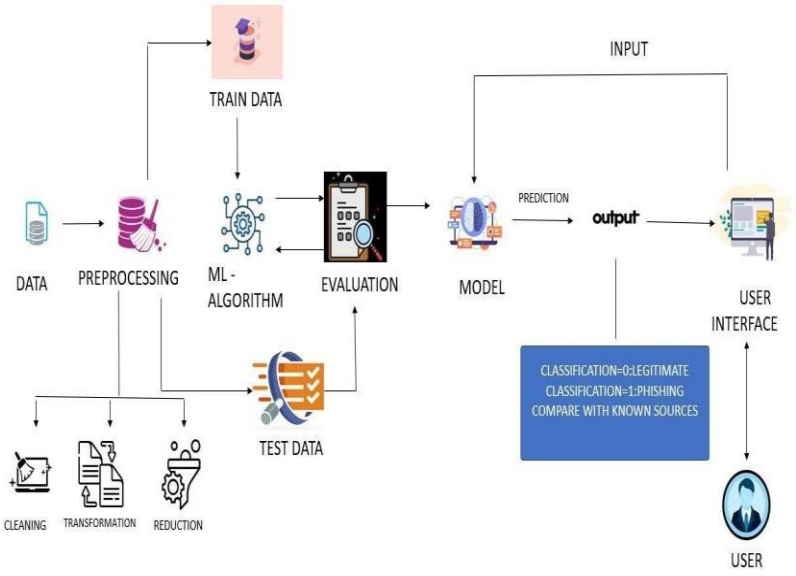

FIGURE 1. Architectural Diagram

4. Methodology

User Interface**:** User can paste link to check that the website is spam or not.
Data Collection:

- The dataset is collected from kaggle dataset.
- The dataset consist with 11057 rows respectively.
- The dataset contains 31 features.

Data Preprocessing: Converting categorical data into numerical data through one-hot encoding. Using imputation the missed values in the dataset are cleaned. Using log transform the skewed data is transformed to normal form.

Model Training: SVM Algorithm: SVM is a Supervised Machine learning algorithm, it requires labelled data to be trained and it will achieve the best separation of data with the boundary around the hyperplane.

Random Forest: Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Decision Tree: Decision tree divide the data set into smaller data sets based on the descriptive features until we reach a small enough set that contains data points that fall under one label.

XG Boost (CatBoost): CatBoost is an algorithm for gradient boosting on decision trees. During training, a set of decision trees is built consecutively. Each successive tree is built with reduced loss compared to the previous trees. The number of trees is controlled by the starting parameters.

Flask Integration: User Interface and our model is been connected by using python flask.

5. Result and Discussion

FIGURE 2. User Interface

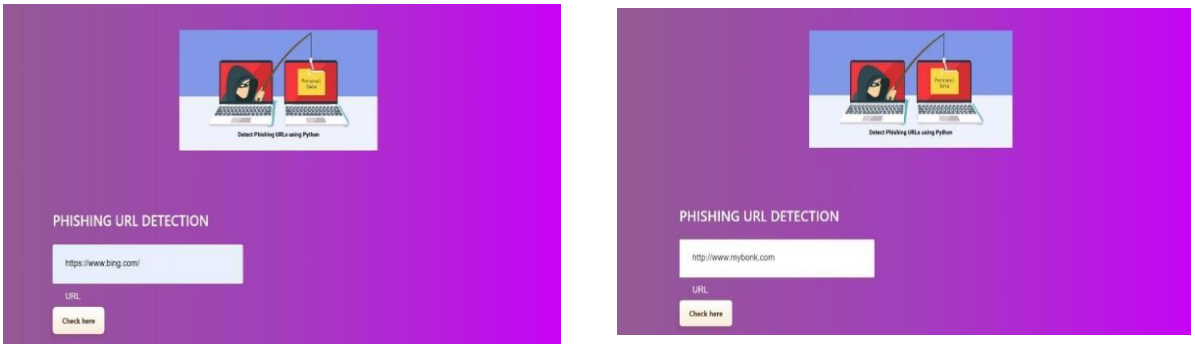

FIGURE 3. Entering the URL

FIGURE 4. Safe site

FIGURE 5. Phishing site

6. Conclusion and Future Scope

It can be concluded with confidence that the cat boost classifier which has high accuracy rate andit makes prediction efficiently. Where the user can enter their URL or link to make prediction as it is a phishing link or not.

Reference

[1]. Anti-phishing Working Group (APWG) Phishing Activity TrendsReport 4th quarter 2020,https://docs.apwg.org/reports/apwg trends report q4 2020.pdf

[2]. FBI Internet Crime Report 2020,https://www.ic3.gov/Media/PDF/AnnualReport/2020 IC3Report.pdf

[3]. Verizon 2020 Data Breach Investigation Report, https://enterprise.verizon.com/resources/reports/2020- data-breachinvestigations- report.pdf

[4]. World Health Organization, Communicating for Health, Cyber Security, https://www.who.int/about/communications/cyber-security

[5]. Ye Cao, Weili Han, and Yueran Le, "Anti-phishing based on automated individual white-list,"Proceedings of the 4th ACM workshop on Digital identity management-DIM 08, pp. 51-60, 2008

[6]. M. Sharifi, and S. H. Siadati, "A phishing sites blacklist generator,2008 IEEE/ACS International Conference on Computer Systems and Applications, pp. 840-843, 2008

[7]. N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection based associative classification data mining," Expert Systems with Applications, vol. 41, no.13, pp. 5948-5959, 2014

[8]. L. Wenyin, G. Huang, L. Xiaoyue, Z. Min, and X. Deng, "Detection of phishing webpages based on visual similarity," Special interest tracks and posters of the 14th international conference on World Wide Web-WWW 05, pp. 1060-1061, 2005

[9]. C. L. Tan, K. L. Chiew et al., "Phishing website detection using url assisted brand name weighting system," 2014 International Symposium on Intelligent Signal Processing and Communication Systems(ISPACS), IEEE, pp. 054-059, 2014

[10]. K. L. Chiew, E. H. Chang, W. K. Tiong et al., "Utilisation of website logo for phishing detection,"Computers & Security, vol. 54, pp. 16-26, 2015

[11]. K. M. kumar, K. Alekhya, "Detecting phishing websites using fuzzy logic," International Journal of Advanced Research in Computer Engineering Technology(IJARCET), vol. 5, no. 10, 2016

[12]. Rishikesh Mahajan, and Irfan Siddavatam, "Phishing website detection using machine learning algorithms," International Journal of Computer Applications(0975-8887), vol. 181, no. 23, 2018

[13]. Jitendra Kumar, A. Santhanavijayan, B. Janet, Balaji Rajendran, and Bindhumadhava BS, "Phishing website classification and detection using machine learning," International Conference on Computer Communication and Informatics(ICCCI), 2020

[14]. Mehmet Korkmaz, Ozgur Koray Sahingoz, Banu Diri, "Detection of phishing websites by using machine learning-based URL analysis,"11nth International Conference on Computing, Communication and Networking Technologies(ICCCNT), 2020

[15]. Mohammad Nazmul Alam, Dhiman Sarma et al., "Phishing attacks detection using machine learning approach," 3rd International Conference on Smart Systems and Inventive Technology(ICSSIT), 2020

[16]. Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi, Touseef J. Chaudhery, "Intelligent phishing website detection using Random Forest classifier," International Conference on Electrical and Computing Technologies and Applications(ICECTA), 2017

[17]. Structure of a URL – image,https://towardsdatascience.com/phishingdomain-detection-with-ml- 5be9c99293e5

[18]. Rami M. Mohammad, Fadi Thabtah, Lee McCluskey, "Phishing websites features"International Journal of Engineering Research & Technology (IJERT)ISSN: 2278-0181Published by, www.ijert.orgNCREIS - 2021 Conference ProceedingsVolume 9, Issue 13