Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 7-Issue 1, January 2024
Paper: 63

Detection of Phishing Attack UsingMachine Learning Techniques

H. Aswini, P. Anuvasanthan, K. Barathraj, V. AjithKumar, J. Pasithahamed
Department of Computer Scienceand Engineering,
IFET College of Engineering Villupuram, Tamil Nadu, aswiniharikrishnan@gma il.com

Abstract— **The chapter discusses a range of research spanning through over the amount of researches carried out on phishing detection and associated work. The chapter is prepared as follows: first and foremost, a rapiddive-in to the which means of phishing in small print to enlighten the reader on why phishing is an essential location of the lookup is given; second, distinct current anti- phishing methods are examined. When a hyper-link is clicked, the customer is redirected to a spoofed website, performing to be from the reputable entity we talk about the techniques used for detection of phishing web sites primarily based on significant properties.**

Keywords—Phishing-Attack, malicious-url, Machine-Learning models, legitimate, Hyperlinks, Detection Techniques

I. INTRODUCTION

Our lives have become more and more dependent on the Internet, but it also provides opportunities for malicious activity such as phishing. Phishers use social engineering or mock websites to deceive individuals and organizations into revealing their account ID, username, and password. Despite the many methods proposed to detect phishing websites, Phishers have evolved their methods to avoid detection. Despite the many methods proposed to detect phishing websites, fisheries have evolved their methods to avoid detection.

Most phishing attacks have some characteristics that can be identified by machine learning. It aims to each folks and organizations, induces them to click on URLs that seem to be secure, and steal personal facts or injects malware on our system. Different computer mastering algorithms are being used for the detection of phishing URLs, that is, to classify a URL as phishing or legitimate.

In this work, we evaluate a goal number of computing devices gaining knowledge of strategies used for this purpose, alongside with datasets and URL aspects usedto educate the machine learning models. Phishing assaults have SaaS and Webmail web sites have been downand assaults on.
Ecommerce websites escalated, whilst assaults on media corporations lowered barely from 12.6% to 11.8%. In mild of the revailing pandemic situation, there have been many phishing assaults that make the most the world focal point onCovid-19. According to WHO, many hackers and cyber scammers are sending fraudulent emails and What Sapp messages to people, taking gain of the covid 19 diseases.

II. LITERATURE REVIEW

In order to forecast an individual's personality, the motive or purpose of phishing is data, cash or private data stealing from the website. The pleasant approach for averting contactwith the phishing internet web page is to realize the actual time with malicious url phishing web sites can be decided on the foundation of their domains they commonly are associated a url which desires to be registered low level area and upper-level domain, path, query

[2] P. BARRACLOUGH, G.DANIEL, N.ASLAM(2015)

Phishing assaults are on the rise, resulting in millions of dollars in losses each year, particularly in online transaction Toolbars and filters that display user warnings against phishing websites have been used in the past to combat phishing attempts. Despite current solutions, there is still an inadequacy in online transaction's due to lack of accuracy in real time solutions. To detect phishing websites and notify users from phishing assaults, and the provide only 100 phishing web sites and 100 non suspicious website. that interface technique is the field of phishing website identification .

[3] ABDULGHANI ALI AHMED, NURUL AMIRAHABDULLAH (2016)

Web spoofing entices customers to join with bogus web sites as a substitute of the authentic ones. The principal intention of this assault is to steal personal facts from users. The attacker develops a 'shadow' internet site that seems to be same to be unique sites this deception permits the attacker to view and edit any facts the sufferer provides. This learn about offers a

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 7-Issue 1, January 2024
Paper: 63

phishing internet site detection approach primarily based on inspecting internet web page Uniform Resource Locators (URLs).The proposed method assessments the url of suspected on-line pages to distinguish between actual and fraudulent internet pages.to notice phishing internet pages, URLs are examined based totally on unique features. The recommended solution's overall performance is assessed by the use of the yahooDataset.

[4] G KUMARI, M NAVEEN KUMAR, A MARYSOWJANAYA (2017)

A large number of people buy things online and pay for them using numerous website Multiple website frequently request sensitive information such as username, password, or credit card information for authentication. However, there are certain phishing websites. Which them uses that information for nefarious purposes. We developed a flexible and successful solution based on data mining concept. But the solution will not provide the best accuracy. We use Regression algorithm. This phishing detection rate, various significant features such as url, domain identify and security can be used to detect website.

III. Existing Work

An existing system the finds the particular kind of urls and the provide 81% accuracy responses. The existing process uses a Bayesian Model to discover the phishing web sites. The classifiers can classify the textual content material and image content, Text classifiers to classify the textual content material and image classifier is to classify the photo content. Bayesian mannequin estimates the threshold value Algorithm combines the each classifier effect and decides whether or not the website online is phishing or not. The performance of existing system will not provide the best accuracy for malicious URL and the display result by using only two methods, either false or true. If its false, the url is bad. if it's true ,the particular is good.

IV. PROPOSED WORK

The proposed system the use nine techniques in machine learning models for detecting the malicious phishing urls. The user will easily identify the given link is malicious or not.

A. Gradient Boosting classifier

A Gradient Boosting is a type of machine learning boosting technique. It builds a better model by merging earlier models until the best model reduces the total prediction error. Also referred to as a statistical forecasting model, the main idea of gradient boosting is to attain a model. In gradient boosting every predictor corrects its predecessor's error. The combined weekend dataset and provides the result.

Y(pred)= y1+(eta * r1) + (eta * r2) + ... + (eta * rN)

B. Cat Boost Classifier

Cat Boost is primarily based on gradient boosted decision trees. During training, a set of choice bushes is developed consecutively. Each successive tree is constructed with decreased loss in contrast to the preceding tree. The variety of trees is managed by means of the beginning parameters.

Cat = CatBoostClassifier(learning_rate = 0,1)Cat-fit(X_train,y_train)

C. XGBoost Classifier

XGBoost is an implementation of gradient boosted decision trees designed for speed performance that is dominative competitive machine learning. In this classifier will provide better accuracy performance

Xgb = XGBClassifier()Xgb.fit (X- train.Y_train)

D. Multi-layer Perceptron Classifier

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 7-Issue 1, January 2024
Paper: 63

MLPClassifier stands for Multi-layer Perceptron classifier which in the identification itself connects to a neural network. Unlike different classification algorithms suchas aid vectors or naïve Bayes classifier, MLPclassifier depends on an underlying neural community to operate the feature of classification

Mlp=MLPClassifier()
Mlp = GridSearchCV(mlpc, parameter_space)Mlp.fit(X_train,Y_train)

E. Logistic Regression

Logistic regression predicts the output of a unique variable. Therefore the consequence has to be specific or discrete value. Logistic Regression is plenty compared to the linear Regression barring that how they are used. Its fixing the classification troubles.

Log= logisticRegression()Log.fit(X_train,Y_train)
y-train_log = log.predict(X_train)

F. K-Nearest Neighbors:Classifier

K-Nearest Neighbor is one of the easiest Machine Learning algorithms primarily based on supervised Learning Techniques. The K-NN algorithm assumes the similarity between the new case/data and on hand instances and put the new case in handy classifier.

Knn = KNeighborsClassifier(n_neighbors=1)Knn.fit(X_train.Y_train)

G. Support vector machine classifier

The Support vector computing device is one of the mostf a m o u s supervised mastering algorithms which is used for classification as nicely as Regression problems. The aim of the svm algorithm is to create the satisfactory line or selection boundary that can furnish the new statistics factor in future.

Param_grid = {'gamma':[0.1],'kernel' :['rbf','linear']} Svc.fit(X_train,y_train)
storeResults('supportvectormachine',acc_test_svc,f1_score_t est_svx,||recall_score_train_svc,precision_score)

H . Naïve Bayes : Classifier

A Naïve Bayes algorithm is a supervised mastering algorithm, which is primarily based on Bayes theorem and used for fixing classification problems. it is primarily used in textual content, image classification that consists of a high-dimensional coachingdataset. Naïve Bayes Classifier is one of the easy and high quality classifications it can make rapid predictions.

Nb = GaussianNB() Nb.fit(X_train,Y_train)
x.(metrics.classification_report(y_test,y_test_Nb))

V. EXPERIMENTS

A. Dataset Information

A series of internet size URLs for 11000 + websites. Each pattern has 30 internet site parameters and a category label figuring out it as a phishing internet site or now not (1 or - 1). The dataset is, contain 11054 samples and 32 features.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 7-Issue 1, January 2024
Paper: 63

B. Experiments

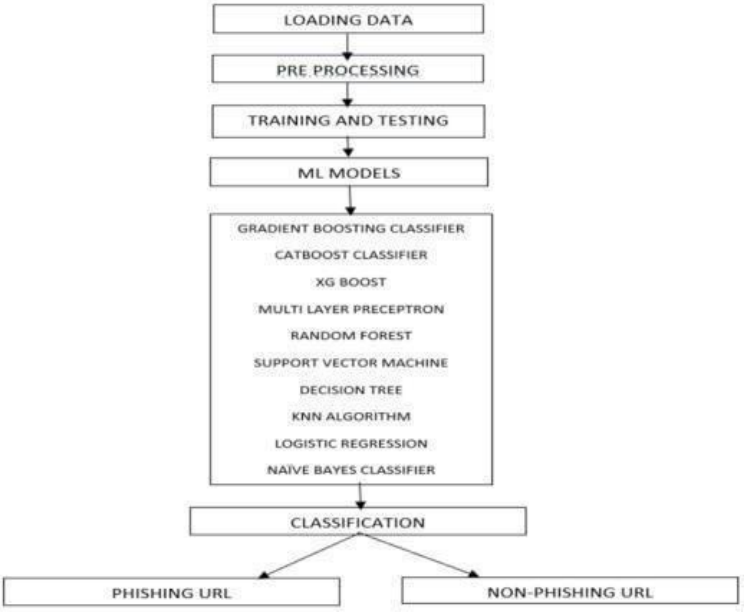


Fig. 1. Architecture diagram of Proposed System

The proposed system the give 11000 websites. The architecture the first load a given Dataset. The data are familiarized with Data & EDA The data frame methods are used to extract the features. The next visualizing the data a plot and graphs are displayed to find how the data is distributed and the how features are related to each other. The splitting the Data into train test sets the split into 80-20 and the evaluate by dependent and independent and the model is building and training by supervised learning techniques our data set comes under the regression problem, as the prediction of malicious link is a continuous number of floating point number in programing terms. The supervised machine learning models (regression) consider training the dataset.

Fig 1. The accuracy of visualize the correlation of Heatmap

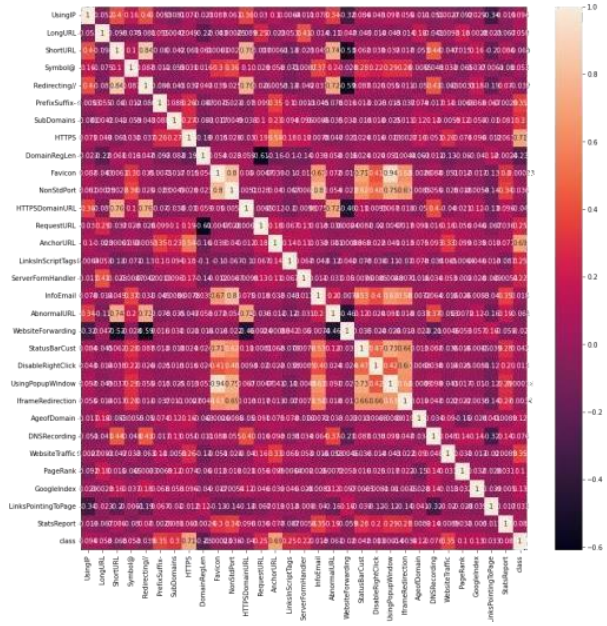


Fig. 2. Accuracy graph of Training and Test model withrespect to number of URLs

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 7-Issue 1, January 2024
Paper: 63

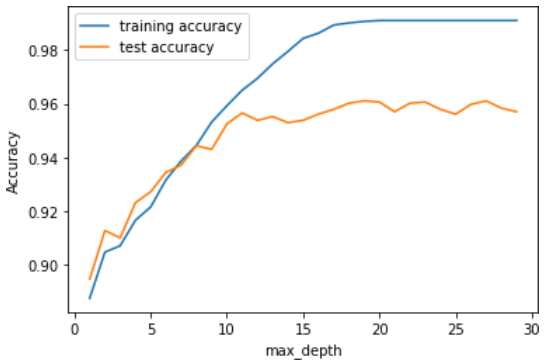
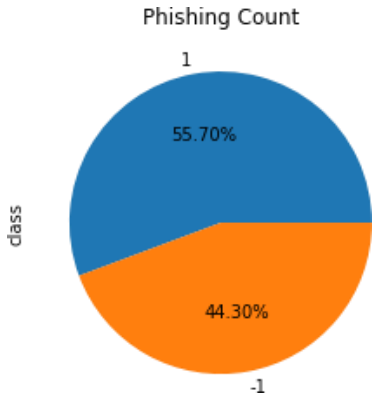


Fig. 3. Pie Chart of phishing count and detection class ofmalicious urls.



C. Performance Evaluation Metrics

A confusion matrix is a precision of number of urls that are actually phishing out of all the urls predicted for phishing it measures the classifiers exact. The formula to calculate precision by the Equation of the matrix.

- True Positive (TP): Real URL are non-malicious and the predicted URL is non-malicious.
- True Negative (TN): Real URL are malicious but the predicted URL is non-malicious.
- False Positive (FP): Real link is malicious but thepredicted link is non-malicious.
- False Negative (FN): Real URL are non-malicious butthe predicted URL is malicious.

TABLE I. CONFUSION MATRIX

	Malicious	**Non-Malicious**
Malicious	TP	FN
Non-Malicious	FP	TN

The quality of the Machine learning algorithms can be scaled using metrics like accuracy, precision etc. Some of themetrics are discussed below:

Accuracy

Accuracy means the amount of correctly categorizedcases.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 7-Issue 1, January 2024
Paper: 63

Accuracy =0.91 or 98 out of 100

Precision

Precision means the true positive to the sum of predictedpositive.

Precision =legitimately_URLS/Hyperlinks

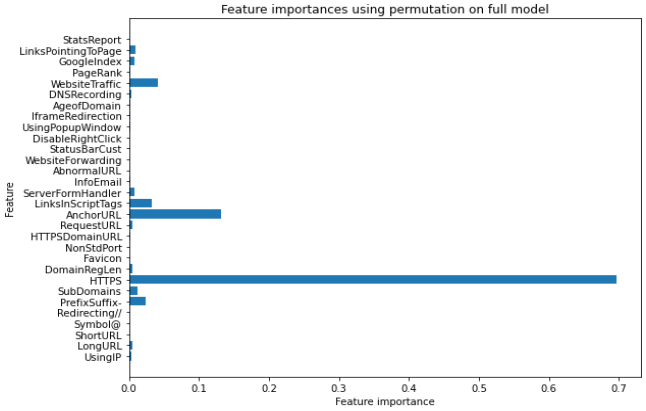


Fig. 4. Performance Metrics of Machine Learning modelsusing Phishing dataset

VI CONCLUSION

The article closing take away structure this assignment is to discover more than a few computer getting to know models, operate Exploratory Data Analysis on phishing dataset and perception their features and creating this pocket book helped me to analyze a lot about the points affecting the fashions to observe whether or not URL is protected or not, Additionally I got here to comprehend how to tune mannequin and how they have an effect on the mannequin performance. He last conclusion on the phishing dataset is that the like “HTTPS”, ”Anchor url”, website Traffic has greater significance to classify URL is phishing url or not.

REFERENCES

1. Security issues, and market interest. In: E-business specialists, commercial center arrangements, security issues, and market interest, London, UK.
2. [Online]. Accessible: http://www. antiphishing.org/assets/apwg- reports/. Gotten to 8Feb2013
3. Kaspersky Lab (2013) Spam in January 2012: love, governmental Issues and game.[Online].
4. Available:http://www.kaspersky.com/about/news/spam/2012
5. Seogod (2011) Black Hat SEO. Search engine optimization Tools.[Online].Accessible:http://www.seobesttools.com/dark cap website optimization/. Gotten to 8 Jan2013
6. Dhamija R, Tygar JD, Hearst M (2006) Why phishing works. In: Proceedings of the SIGCHI meeting on human factors in figuring frameworks, Cosmopolitan Montre 'al,Canada
7. Cranor LF (2008) A system for thinking about the human tuned in. In: UPSEC'08
8. Proceedings of the first meeting on ease of use, brain science, and security, Berkeley,CA,USA
9. Miyamoto D, Hazeyama H, Kadobayashi Y (2008) An assessment of AI based techniquesfor recognition of phishing destinations. Aust J Intell Inf Process Syst 10(2):54–6