INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 4 - V7I4-1324)
Available online at: https://www.ijariit.com

Phishing website detection using Machine Learning

R. Dhatri
raavidhatri@gmail.com
Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh

N. Shafiyabi
nadigaddashafi123@gmail.com
Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh

U. Nithish
nithish.rana7690@gmail.com
Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh

P. Vamsi
vamsichowdary443@gmail.com
Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh

K. Subrahmanya Kousik
kousikkappaganthu@gmail.com
Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh

ABSTRACT

Today internet or websites plays a major role in every person life. Almost Every sector using the website services like banking sector using some apps for payment, passport sector is using for registration and renewal of passport, government sector for application of PAN cards, Adhar cards etc. This makes the human life simpler. But this provides an opportunity to make phishing websites. The phishing websites is same as legitimate or real websites. The purpose of the phishers is to stole their personal information, account ID, Passwords from individuals and organizations. They even add some tricks by asking security questions like pet name, city name to gain user's trust. Although legitimate and phishing looks like same there are some features that make difference between those two things. That features such as IP address, URL length, having @ symbol, double slash redirection, Prefi and suffix, having subdomain, domain registration link, HTTPS tokens, Request URL, URL of anchor, disabling right click, using pop-up window and some more. Already many approaches are proposed for detecting phishing websites machine learning is most appropriate one. This is because there are some common features which can be identified by machine learning. In this paper we used random forest algorithm to detect phishing websites based on feature that make difference between both of the websites.

Keywords: *Phishing Websites, Random Forest, Machine Learning, NumPy, Pandas.*

1. INTRODUCTION

Internet services is used by each and every person and it becomes the important part in our life. Every work become simpler through this network. But the user may not aware of phishing websites as they are much similar to legitimate websites. The Phishers develop the phishing websites to improve their business needs by gathering our personal information like passwords, username, Bank details. That information is misused and results to loss of finance or money. This phishing websites is increasing day by day. So, detecting Websites is very important thing. Although the Phishing websites and legitimate websites looks similar there are some features that differentiate between them . The machine Learning is a powerful tool for detecting patterns in data and methods have made it possible to detect the common phishing traits. In this paper we studied about Random Forest algorithm which evolved form Decision Tree algorithm that extracts the features and apply preprocessing to detect phishing websites.

2. LITERATURE REVIEW

Machine Learning(ML): A study of developing computer algorithms for transforming data into intelligent action is known as "Machine Learning". It originates at an intersection of statistics, data science and computer science. It is seen as a part of Artificial Intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as Credit Card Fraud Detection, Emotion Detection, Sentimental Analysis, email filtering and many more, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights

subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data scientists will increase, requiring them to assist in the identification of the most relevant business questions and subsequently the data to answer them. The Learning system of ML is divided into three parts that is a decision Process, which In general, machine learning algorithms are used to make a prediction or classification. Based on some input data, which can be labelled or unlabeled, your algorithm will produce an estimate about a pattern in the data. Next is an Error Function, which means an error function serves to evaluate the prediction of the model. If there are known examples, an error function can make a comparison to assess the accuracy of the model. Last one is a model optimization Process, which means if the model can fit better to the data points in the training set, then weights are adjusted to reduce the discrepancy between the known example and the model estimate. The algorithm will repeat this evaluate and optimize process, updating weights autonomously until a threshold of accuracy has been met.

3. DETECTING PHISHING WEBSITES USING RANDOM FOREST ALGORITHM

Random forest algorithm is one of the most powerful algorithms in machine learning technology and it is based on concept of decision tree algorithm. Random forest algorithm creates the forest with number of decision trees. High number of trees gives high detection accuracy. Creation of trees are based on bootstrap method. In bootstrap method features and samples of dataset are randomly selected with replacement to construct single tree. Among randomly selected features, random forest algorithm will choose best splitter for the classification and like decision tree algorithm; Random Forest algorithm also uses gini index and information gain methods to find the best splitter. This process will get continue until random forest creates n number of trees. Each tree in forest predicts the target value and then algorithm will calculate the votes for each predicted target. Finally random forest algorithm considers high voted predicted target as a final prediction. Decision trees are non-parametric classifiers. As its name indicates, a decision tree is a tree structure, where each nonterminal node denotes a test on an attribute, each branch represents an outcome of the test, and the leaf nodes denote classes. The basic algorithm for decision tree induction is a greedy algorithm that constructs the decision tree in top-down recursive divide-and-conquer manner . At each non-terminal node, one of attributes is chosen for the split. The attribute that gives the maximum information gain is chosen for the split. A well-known algorithm for decision trees is the C4.5 algorithm where entropy is used as a criterion to calculate the information gain. The information gain is defined as the difference between the entropy before the split and the entropy after the split.

4. METHODOLOGY

4.1 Tools used

Python and Open NumPy, Panda's libraries are the main tools. Python is a high-level programming language in which the system models are defined. As a result, you'll need a fully functional Python 3.5+ environment including the sklearn, numpy, and pandas' libraries. NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Numeric, the ancestor of NumPy, was developed by Jim Hugunin. Another package Numarray was also developed, having some additional functionalities. In 2005, Travis Oliphant created NumPy package by incorporating the features of Numarray into Numeric package. There are many contributors to this open-source project. Using NumPy, mathematical and logical operations on arrays can be performed. This tutorial explains the basics of NumPy such as its architecture and environment. It also discusses the various array functions, types of indexing, etc. Using NumPy a developer can perform following operations. First, Mathematical and logical operations on arrays. Second is Fourier transforms and routines for shape manipulation. Third, Operations related to linear algebra. NumPy has in-built functions for linear algebra and random number generation.

Pandas, Pandas is an open-source python package built on top of Numpy developed by Wes McKinney**.** It is used as one of the most important data cleaning and analysis tool. It provides fast, flexible, and expressive data structures. To install Pandas, use command – pip install pandas.

4.2 Process

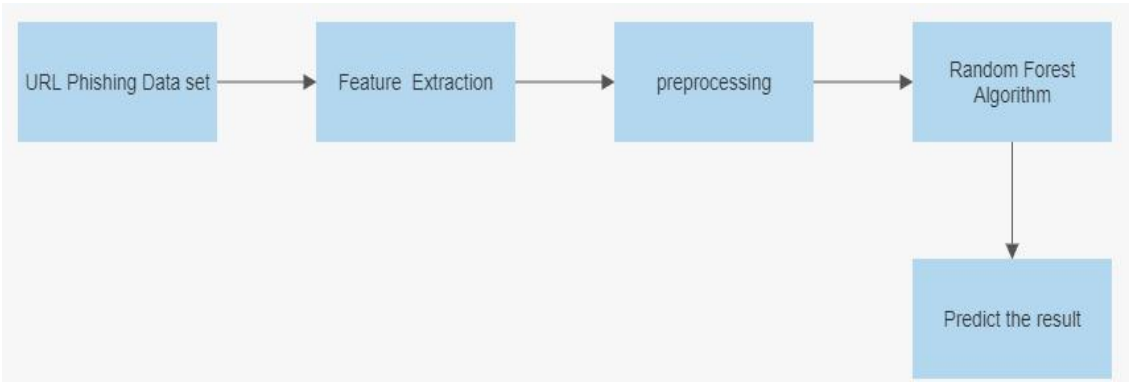
In this project we use random forest algorithm to detect phishing websites. We are building a model in such a way that the model comprises of feature extraction from sites and classification of website. In feature extraction, 30 features have been taken to differentiate real websites with phishing websites.
That feature includes URL length, Request URL, having IP address in the URL etc.
This feature is useful to know whether the website is phishing or useful.
These three different types of features are used to detect phishing websites by applying machine learning algorithms.

(a) URL and Derived Features: These features are included in the address of the website Phishers generally adopt some methods to attack like providing in instead of URL.
(b) Pages source code based Features: A common trick employed by phishers is to make the interface textually and graphically very similar to a legitimate webpage.
(c) Domain based Features: After splitting the dataset, we shortlisted the different features for the classification of URLs.

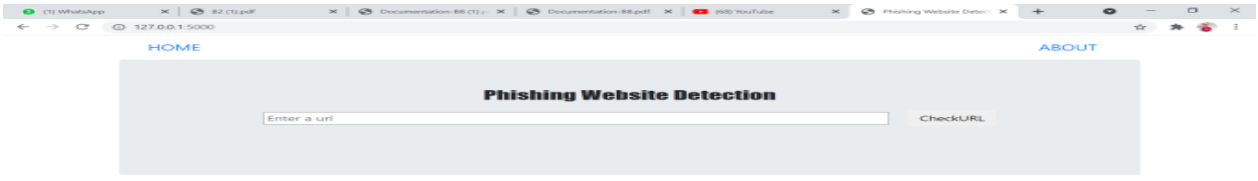
For testing the results obtained, we used 3 parameters:
Accuracy, Recall and False Positive Rate (FPR).

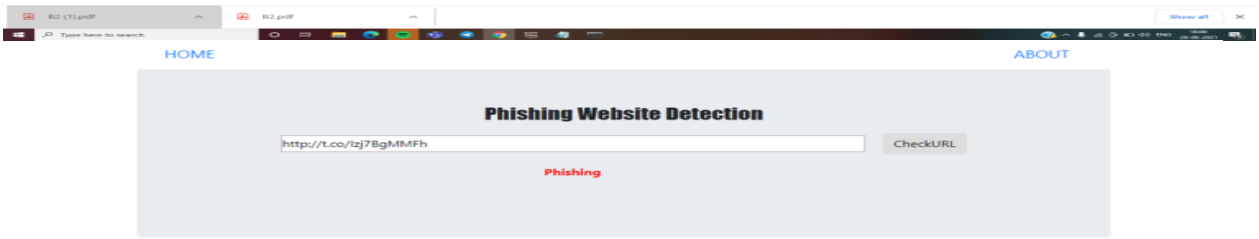
(a) Accuracy: It is the ratio of number of correct predictions to the total number of input samples.
(b) Recall: It is the ratio of number of true positives to the total number of predicted positives.
(c) False Positive Rate (FPR): It is the ratio of number of samples incorrectly identified as positive to total number of actually negative samples.

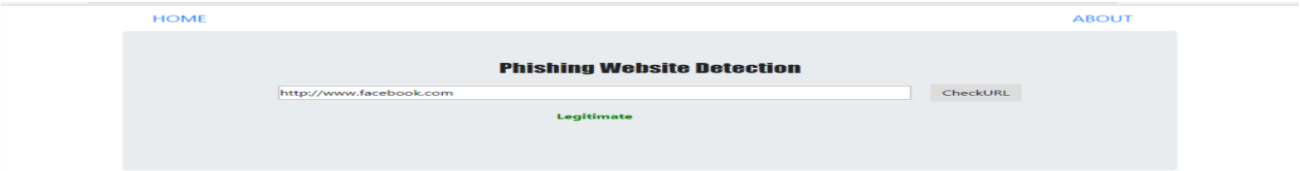
S. No	Test Scenario(Input given in the User Interface)	Test output
1.	http://133.130.103.10/1/	Phishing
2.	http://www.livinglegendsltd.com/	Legitimate
3.	http://verification-mobile-nab.com/cgi/e8e11a52ed442dc4b327eaa2f19c2521/login/	Phishing
4.	http://umeda.com.br/bba/BOA/home/	Phishing
5.	http://www.austinchronicle.com/issues/dispatch/1999-12-10/music_feature3.html	Legitimate
6.	http://ginatringali.com//al/alibaba21012015/alibaba21012015/666/index.html	Phishing
7.	http://www.awn.com/mag/issue5.03/5.03pages/evanierforay.php3	Legitimate



Steps to implement the algorithm:
Step 1:-Data Collection.
Step 2:-Data Preprocessing.
Step 3:-Feature Extraction.
Step 4:- Apply Random Forest Algorithm.
S**tep 5:-** Predict the result.
Result and Discussion:
The proposed model has been developed to recognize Face.







5. CONCLUSION

This is to enhance detection method to detect phishing websites using machine learning technology. We achieved better detection accuracy using random forest 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 random forest algorithm of machine.

6. REFERENCES

[1].https://digitalguardian.com/blog/whatphishing-attack-defining-and-identifying-different-types-phishingattacks
[2] https://ieeexplore.ieee.org/abstract/document/87695714
[3]https://resources.infosecinstitute.com/category/enterprise/phishing/the-phishing-landscape/phishing-data-attack-statistics/#gref
[4]http://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-work/

