A

REPORT

ON

Detection of Phishing Websites using Machine Learning

Submitted By:

SUPERVISED BY

Table of Content

Introduction

What is Phishing?

Phishing refers to luring techniques used by identity thieves to fish for personal information in a pond of unsuspecting Internet users. It is a general term for the creation and use by criminals of e-mails and websites that have been designed to look like they come from well-known, legitimate and trusted businesses, financial institutions and government agencies. These criminals deceive Internet users into disclosing their bank and financial information or other personal data such as usernames and passwords.

How Is Phishing Committed?

In a typical phishing scheme, criminals who want to obtain personal data from people online first create unauthorized replicas of (or "spoof") a real website and email, usually from a financial institution or another company that deals with financial information, such as an online merchant. The email will be created in the style of emails by a legitimate company or agency, using its logos and slogans. The nature and format of the principal website creation language, Hypertext Markup Language, make it very easy to copy images or even an entire website. While this ease of website creation is one of the reasons that the Internet has grown so rapidly as a communications medium, it also permits the abuse of trademarks, trade names, and other corporate identifiers upon which consumers have come to rely as mechanisms for authentication.

Nowadays it has become a main area of concern for security researchers because it's easy to create a fake website by using HTML code of the legitimate website.Experts can identify a fake website but not an ordinary user and that user becomes victim of Phishing attack. It is very important to enhance Phishing Detection Techniques.

There are many methods to detect a Phishing website like Blacklist method, Heuristic based detection method and Machine Learning Methods.We will only deal with Machine Learning Technology for detection of Phishing URLs by extracting and analyzing various features of legitimate and Phishing URLs.Decision Tree, Random Forest and Support Vector Machine are used to detect Phishing Websites.

Objective:

The objective of this report is to identify Phishing websites using different ML methods and to increase the accuracy in detection.

Dataset

Sub Description

Name-Phishing Website Data Set

Topics-cyber crime, phishing

Short Description

In this dataset, light is shed on the important features that have proved to be sound and effective in predicting phishing websites.

Long Description

Although many articles about predicting phishing websites have been disseminated, no reliable training dataset has been previously published publically, maybe because there is no agreement in literature on the definitive features that characterize phishing webpages, hence it is difficult to shape a dataset that covers all possible features.This dataset collected mainly from: PhishTank archive, MillerSmiles archive, Googles searching operators.

Data Set Characteristics:

Number of Instances:2456

Area:Computer Security

Attribute Characteristics:Integer

Number of Attributes:30

Sub Dataset Link

https://archive.ics.uci.edu/ml/datasets/phishing%20websites

Machine Learning Algorithms

We will use these three Machine Learning Classification models to detect Phishing Websites:

1.Decision Tree Algorithm

It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

Creation of a Tree

A tree is built by splitting the source set, constituting the root node of the tree, into subsets—which constitute the successor children. The splitting is based on a set of splitting rules based on classification features.This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same values of the target variable, or when splitting no longer adds value to the predictions. This process of top-down induction of decision trees (TDIDT) is an example of a greedy algorithm, and it is by far the most common strategy for learning decision trees from data.

In our project we are going to use classification tree as our predicted outcome is the class to which the data belongs.

Advantages

1.Simple to understand and interpret.

2.Able to handle both numerical and categorical data.

3.Performs well with large datasets.

4.In built feature selection-It means that the features on top are the most informative.

Limitations

1.A small change in the training data can result in a large change in the tree and consequently the final predictions.

2.Decision-tree learners can create over-complex trees that do not generalize well from the training data.

3.For data including categorical variables with different numbers of levels, information gain in decision trees is biased in favor of attributes with more levels.

2.Random Forest Algorithm

Random forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.Random decision forests correct for decision trees' habit of overfitting to their training set.Rather than just simply averaging the prediction of trees (which we could call a “forest”), this model uses two key concepts that gives it the name random:

1.Random sampling of training data points when building trees

2.Random subsets of features considered when splitting nodes

Construction of a random forest

1.If there are N samples, there are N samples randomly selected (each time one sample is randomly selected, and then returned to continue selection). This selected N samples are used to train a decision tree as a sample at the root of the decision tree.

2.When each sample has M attributes, when each node of the decision tree needs to be split, m attributes are randomly selected from the M attributes, satisfying the condition m << M. Then use some strategy (such as information gain) from these m attributes to select 1 attributes as the split attribute of the node.

3.In the decision tree formation process, each node must be split according to the step 2 (it is easy to understand that if the next attribute selected by the node is the attribute that was used just when its parent node was split, the node has reached the leaf. Node, no need to continue to split). Until it can't be split again. Note that no pruning is done during the formation of the entire decision tree.

4.According to the steps 1~3, a large number of decision trees are created, which constitutes a random forest.

Advantages

1.It can come out with very high dimensional data, and no need to reduce dimension, no need to make feature selection

2.It can judge the importance of the feature

Can judge the interaction between different features

3.Not easy to overfit

4.Training speed is faster, easy to make parallel method

5.It is relatively simple to implement

6.For unbalanced data sets, it balances the error.

If a large part of the features are lost, accuracy can still be maintained.

Limitations

1.Random forests have been shown to fit over certain noisy classification or regression problems.

2.For data with different values, attributes with more values will have a greater impact on random forests, so the attribute weights generated by random forests on such data are not credible.

3.Support Vector Machine Algorithm

In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. The Support Vector Machine (SVM) algorithm is a popular machine learning tool that offers solutions for both classification and regression problems.

Implementation of SVM in Python

In Python, scikit-learn is a widely used library for implementing machine learning algorithms. SVM is also available in the scikit-learn library and we follow the same structure for using it(Import library, object creation, fitting model and prediction).

Advantages

1. It works really well with a clear margin of separation.

2. It is effective in high dimensional spaces.

It is effective in cases where the number of dimensions is greater than the number of samples.

3. It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

Limitations

1.It doesn’t perform well when we have large data set because the required training time is higher.

2.It also doesn’t perform very well, when the data set has more noise i.e. target classes are overlapping.

3.SVM doesn’t directly provide probability estimates, these are calculated using an expensive five-fold cross-validation. It is included in the related SVC method of Python scikit-learn library.

Implementation

Sci-Kit Learn tool will be used to import Machine Learning Algorithms. Dataset will be divided into training set and testing set in different ratios respectively. Each classifier will be trained using training set and testing set. Performance of Classifiers will be evaluated by calculating classifier's accuracy score, false negative rate and false positive rate.

Conclusion

We aim to enhance detection method to detect Phishing websites using ML technology and to achieve maximum accuracy.