Detection of Phishing websites

Faizan Shakeel [KZ4RE0]

Faculty of Informatics, Department of Data Science and Engineering, ELTE-Eötvös Loránd University, Pázmány Péter sétány 1117, Budapest, Hungary

erfury@yahoo.com

Abstract. Website Phishing allows the user to interact with the fake websites rather than the legitimate ones. The main objective of this attack is to steal sensitive information from the users. The attacker creates a website that looks similar to the legitimate website. This allows the attacker to steal and modify any information from the user. In this paper, we trained different machine learning models for detecting phishing websites. The dataset includes 30 phishing website features that make our model more accurate and precise. This project demonstrates an analytical approach for the prediction of phishing websites using a probabilistic model based on SVM, Random forest, and decision tree. The technique used for classification and prediction is based on recognizing typical and diagnostically most important features. These features are provided as input to the classification model for prediction and qualitative analysis.

Keywords: Phishing Detection, Phishing Website, Phishing Attacks

1 Introduction

In this growing world of technology, the internet played a vital role in transferring information from one place to another. In this competitive environment, everyone is trying to move their business online and have a website for their company for client interaction. It has become a valuable and convenient mechanism for supporting public transactions such as e-banking and e-commerce transactions. That has led the users to trust it is convenient to provide their private information to the Internet. As a result, the security thieves that have started to target this information have become a major security problem The plethora of information on the Internet enables businesses to learn about trends that may affect them, observe consumer behavior, discover products that could enhance their service or business and increase their knowledge of the industry. Information that used to be reserved for the most influential people in the industry or academics is now easier to find. As now it gives the scammers a way to steal useful information from users by tricking them. Phishing is a technique where the attacker lures the user to interact with the fake websites rather than the legitimate. Phishing website is one of the internet security problems that target human vulnerabilities rather than software vulnerabilities.

Statistics have shown that the number of phishing attacks keeps increasing, which presents a security risk to the user information according to the Anti-Phishing Working Group (APWG) [1] and recorded phishing attacks by Kaspersky Lab [2], which stated that it has increased by 47.48% from all of the phishing attacks that have been detected during 2016. Recently, there have been several studies that tried to solve the phishing problem. Some researchers used the URL and compared it with existing blacklists that contain lists of malicious websites, which they have been creating, and there are others that have used the URL in an opposite manner, namely comparing the URL with a whitelist of legitimate websites. Additionally, measuring website traffic using Alexa is another way that has been implemented by researchers to detect phishing websites [3].

2 Related Work

2.1 Content-Based Approach

Zhang et al. [4] presented the design and evaluation of CANTINA, a novel, content-based approach to detecting phishing websites, using the well-known TF-IDF algorithm. It analyses the text-based content of a page by itself. They experimented with some simple heuristics that can be applied to reduce false positives. As a result, a pure TF-IDF approach can catch about 97% of phishing sites with about 6% false positives, and with heuristics, it catches about 90% of phishing sites with only 1% false positives. Rao and Ali [5] implemented a desktop application to detect phishing websites using a novel heuristic based on URLs and website content. With the application called Phish Shield, they used copyright, null footer links, zero links of the body HTML, links with maximum frequency domains, and whitelists to detect phishing websites. It achieved an accuracy of 96.57% with a FP of 0.035%.

2.2 URL Approach

A new approach that Nguyen et al. [6] had proposed to detect phishing sites is by deriving different components from the URL and computing a metric for each component. Then, the page ranking will be combined with the achieved metrics to decide whether the websites are phishing websites. The results showed that the technique can detect over 97% of phishing websites. Jeeva and Rajsingh [7] presented a system for prediction phishing URLs by generating rules of association rule mining. They used the apriori algorithm to pick known information from the frequent itemset properties that were extracted from the dataset. Jeeva and Rajsingh [7] also used another algorithm that performs on hidden data to obtain the accuracy of association rules, which is a predictive apriori that engages the confidence and the support techniques that are measured in its accuracy, unlike a priori, which only mark rules that have the confidence technique. As a result, they presented significant features of the URL that distinguish if it is phishing or legitimate.

3 Dataset

The dataset used is publicly available on the UCI Machine Learning Repository which contains the data of more the 10000 websites both legitimate and fake websites. The dataset has total 30 attributes and 11055 row entries.

S.no	Attribute Name	Values
1	Having_IP_Address	{-1, 1}
2	URL_Length	{1, 0, -1}
3	Shortining_Service	{1, -1}
4	Having_AT_Symbol	{1, -1}
5	Double_Slash_Redirecting	{1, -1}
6	Prefix_Suffix	{1, -1}
7	Having_Sub_Domain	{1, -1, 0}
8	SSLfinal_state	{-1, 1, 0}
9	Domain_Registration_length	{-1, 1}
10	Favicon	{-1, 1}
11	Port	{-1, 1}
12	HTTPS_Token	{-1, 1}
13	Request_URL	{-1, 1}
14	URL_of_Anchor	{-1, 0, 1}
15	Links_in_Tags	{-1, 1, 0}
16	SFH	{-1, 1, 0}
17	Submitting_to_email	{-1, 1}
18	Abmormal_URL	{-1, 1}
19	Redirect	{0, 1}
20	On_mouseover	{-1, 1}
21	RightClick	{-1, 1}
22	PopUpWindow	{-1, 1}
23	Iframe	{-1, 1}
24	Age_of_domain	{-1, 1}

25	DNSRecord	{-1, 1}
26	Web_Traffic	{-1, 0, 1}
27	Page_Rank	{-1, 1}
28	Google_Index	{-1, 1}
29	Links_pointing_to_page	{1, 0, -1}
30	Statistical_report	{-1, 1}

Table 1: Attributes and values of the dataset

4 Data preprocessing

Since we want reproducible results, it is a good idea to pre-process the data in advance. The objective is to turn raw data into a form that is convenient for building a model. We do this by splitting the data into a training set (which is used to train the model) and a testing set (which is used to evaluate the performance of the trained model). The pre-processing procedure followed in this example consists of the following steps

1. Import data from the raw csv file
2. Explore its contents
3. Check for missing values and decide how to deal with them
4. Finding the correlations

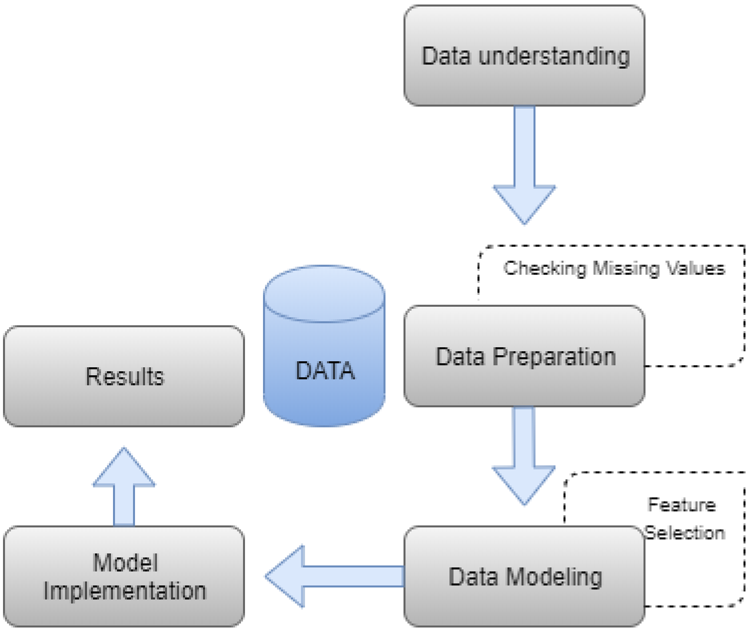


Fig. 1. Data Preprocessing

4.1 Data Cleaning

Data cleaning is the process of changing data in the storage source to make it accurate and correct. Most of the data cleaning software relies on a review of the data set and the way they are attached to the data storage technology. Data cleaning is also known as data scrubbing. Data cleaning is compared to data purging, in which old or useless data is deleted from a data set. Data cleaning involves deleting of the old, incomplete and duplicate data. A data cleansing method uses methods to get rid of syntax errors, typographical errors or fragments of records.

Data Cleaning techniques are applied to remove noise and correct instances in the data. For example, data cleaning can involve transformations to correct wrong data, such as by transforming all entries for a date field to a common format. It can be associated with Missing data and Noisy data.

In case of any missing data, that particular data entry can be:

1. Ignored
2. Manually entered
3. Given the value of a constant
4. Given the value of the mean

The main problem in this work we have shown is that we compare various machine learning models on the basis of their parameters like accuracy. The main part of the project involves to find out the appropriate and suitable algorithm for the efficient diagnosis and classification of the phishing website dataset. The algorithms use mathematical techniques to separate different data features.

5 Models

5.1 SVM

Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. In 1960s, SVMs were first introduced but later they got refined in 1990. SVMs have their unique way of implementation as compared to other machine learning algorithms. Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables.

The Support Vector Machine (SVM) model performs the following results with the dataset provided to it. Accuracy of this model comes out to be about 93 %

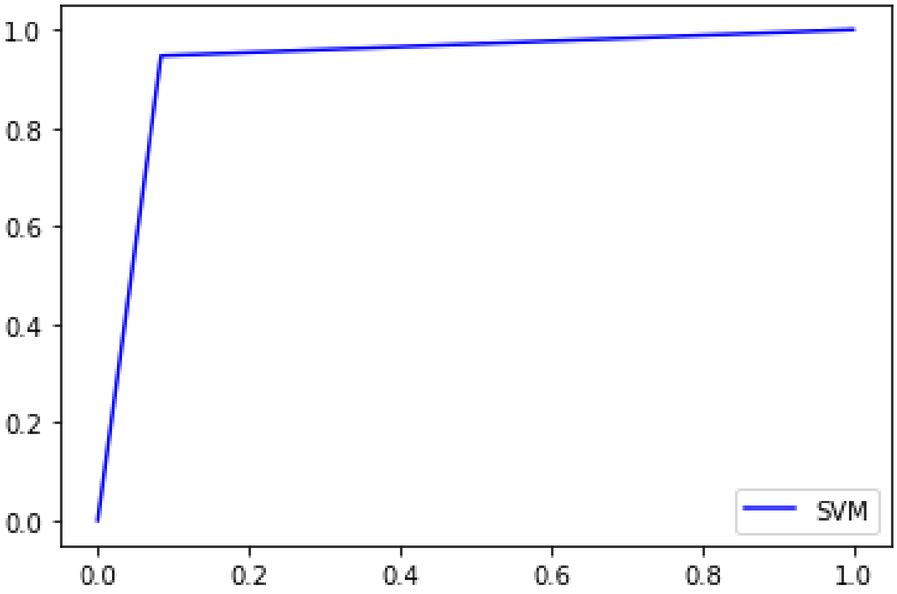


Fig. 2. ROC curve for SVM

5.2 Random Forest

Random forest is a supervised learning algorithm that is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees mean more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting.

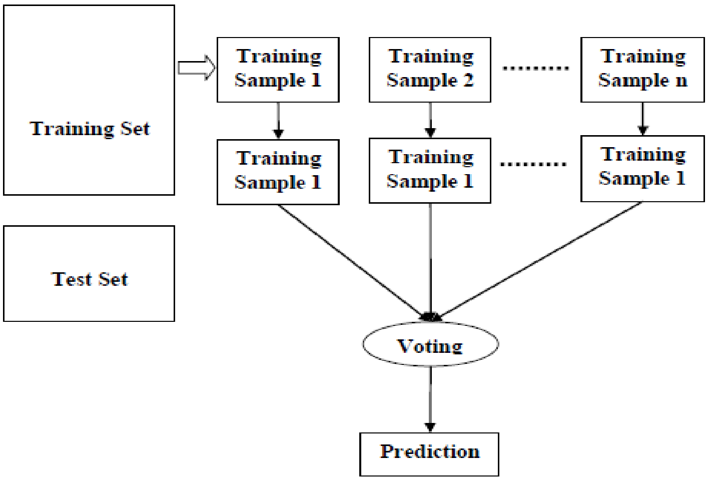


Fig. 3. Random Forest Model

Visualization of a Confusion matrix.

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data.

True Negatives - All samples that were identified as negative labels and were truly negative

False Negatives - All samples that were identified as negative labels and were in fact positive

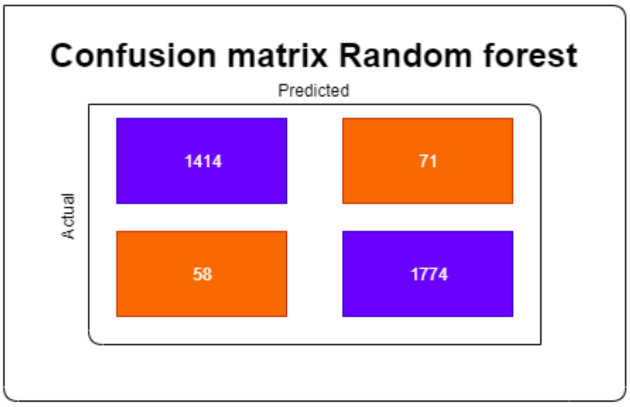
True Positives - All samples that were identified as positive labels and were truly positive

False Positives - All samples that were identified as positive labels and were in fact negative

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\mathit{Accuracy} = \frac{TP + TN}{\mathit{Total}}$$



1414	71
58	1774

Fig. 4. Confusion matrix for Random forest

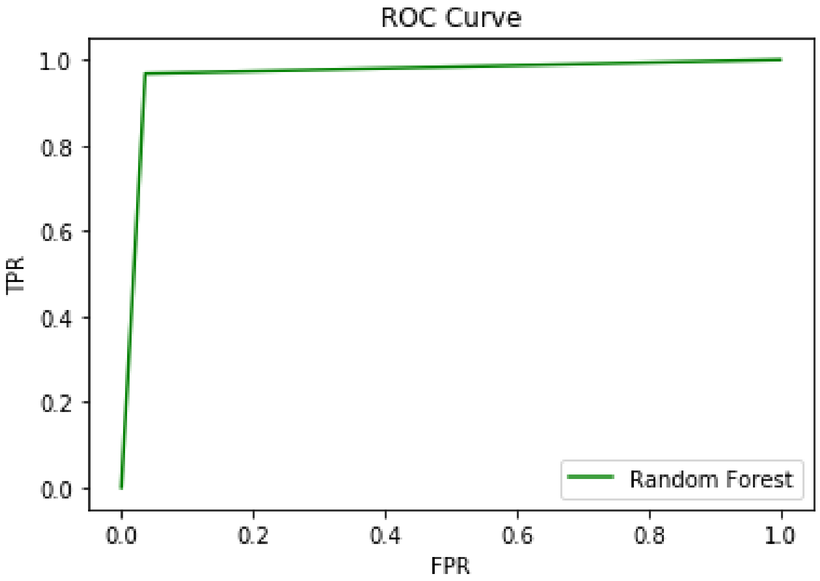


Fig. 5. Roc curve for Random Forest

```
Accuracy with RF classifier: 0.96
```

5.3 Decision Tree

In general, Decision tree analysis is a predictive modelling tool that can be applied across many areas. Decision trees can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions. Decisions trees are the most powerful algorithms that falls under the category of supervised algorithms. They can be used for both classification and regression

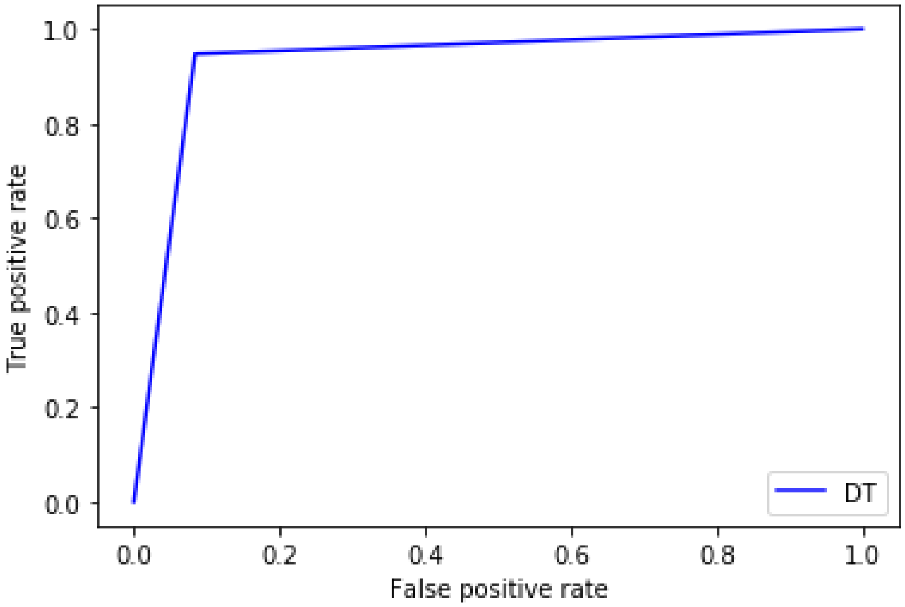


Fig. 6. Roc curve for Decision Tree

```
Accuracy: 0.95
```

6 Model Comparison

As we have used the SVM, Random Forest and decision tree algorithms to detect

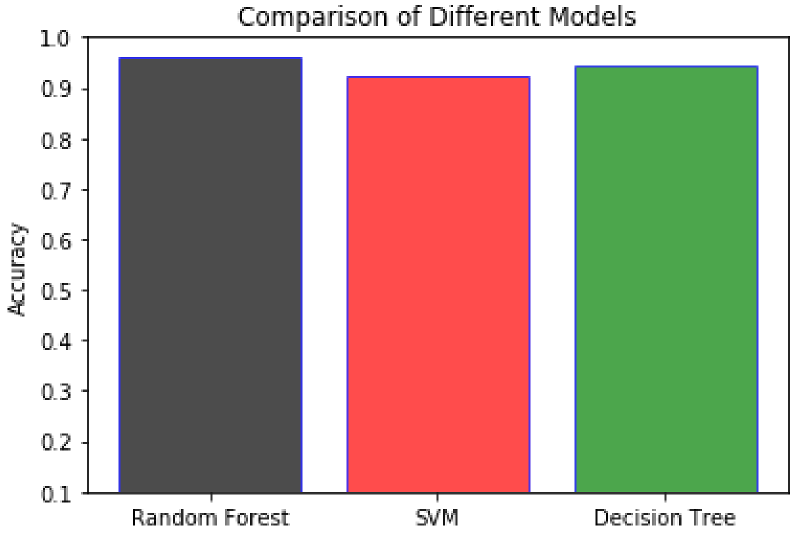


Fig. 7. Comparison of different models used

whether the website is legitimate or fake. The model with the highest accuracy is Random forest.

S.no	**Model**	**Accuracy**
01	SVM	0.93
02	Random Forest	0.96
04	Decision Tree	0.95

Table. 2. Model Accuracy

7 Conclusion

In our project, our goal was to build machine learning algorithms which perform the best accurate results with the phishing website dataset of about 11050 websites. Various models like SVM, Random forest and decision tree have their different range of predictions of the phishing and legitimate website. The random forest gives the best result with an accuracy of 96%.

References

1. AO Kaspersky lab. (2017). The Dangers of Phishing: Help employees avoid the lure of cybercrime. [Online] Available: https://go.kaspersky.com/Dangers-Phishing-Landing-Page-Soc.html [Oct 30, 2017].
2. ” Financial threats in 2016: Every Second Phishing Attack Aims to Steal Your Money” Internet: https://www.kaspersky.com/about/pressreleases/2017 financial-threats-in-2016. Feb 22, 2017 [Oct 30, 2017].
3. L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, ”Detecting phishing web sites: A heuristic URL-based approach,” in 2013 International Conference on Advanced Technologies for Communications (ATC 2013), 2013, pp. 597-602.

4. Y. Zhang, J. I. Hong, and L. F. Cranor, ”Cantina: A Content-based Approach to Detecting Phishing Web Sites,” New York, NY, USA, 2007, pp. 639-648
5. R. S. Rao and S. T. Ali, ”PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach,” Procedia Computer Science, vol. 54, no. Supplement C, pp. 147-156, 2015.
6. L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, ”Detecting phishing web sites: A heuristic URL-based approach,” in 2013 International Conference on Advanced Technologies for Communications (ATC 2013), 2013, pp. 597-602.
7. Z. Zhang, Q. He, and B. Wang, ”A Novel Multi-Layer Heuristic Model for Anti-Phishing,” New York, NY, USA, 2017, p. 21:1-21:6.