Website Phishing ML Project

**Group Members**

Dino Anastasopoulos: 1900661

Timothy Walters: 1855167

Razeen Gani: 1842173

Reece James Peters: 1924514

**Dataset Information**

We chose to study a dataset about Website Phishing.

The dataset consists of 9 attributes, 3 class types and 1353 datapoints. The goal is to predict the classification, that is whether the website is legitimate, suspicious or phishy.

The attributes and their corresponding values are as follows:

(Note: 1 = Legit, 0 = Suspicious, -1 = Phishy)

* **Server Form Handler:**

Once the user submitted his information; the webpage will transfer the information to a server so that it can process it. Normally, the information is processed from the same domain where the webpage is being loaded. Phishers resort to make the server form handler either empty or the information is transferred to somewhere different than the legitimate domain.

* + 1: otherwise
  + 0: SFH redirects to different domain
  + -1: SFH is ‘about : blank’ or empty
* **PopUp Window:**

Usually authenticated sites do not ask users to submit their credentials via a popup window.

* + 1: otherwise
  + 0: rightClick alert showing
  + -1: rightClick disabled
* **Fake HTTPs protocol/SSL final:**

The existence of HTTPs protocol every time sensitive information is being transferred reflects that the user is certainly connected with an honest website. However, phishers may use a fake HTTPs protocol so that users may be deceived, so it is recommended to check that the HTTPs protocol is offered by a trusted issuer.

* + 1: use of https and trusted issuer and age ≥ 2 years
  + 0: using https and issuer is not trusted
  + -1: otherwise
* **Request URL:**

A webpage usually consists of text and some objects such as images and videos. Typically, these objects are loaded into the webpage from the same server of the webpage. If the objects are loaded from a domain other than the one typed in the URL address bar, the webpage is potentially suspicious.

* + 1: request URL < 22%
  + 0: 22% ≤ request URL < 61%
  + -1: otherwise
* **URL of Anchor:**

Similar to the URL feature, but here the links within the webpage may point to a domain different from the domain typed in the URL address bar.

* + 1: URL anchor % < 31%
  + 0: 31% ≤ URL anchor < 67%
  + -1: otherwise
* **Web Traffic:**

Legitimate websites usually have high traffic since they are being visited regularly. Since phishing websites normally have a relatively short life; they have low web traffic.

* + 1: Web Traffic > 150 000
  + 0: 150 000 ≥ Web Traffic ≥ 50 000
  + -1: otherwise
* **URL Length:**

Phishers hide the suspicious part of the URL to redirect the information submitted by users or redirect the uploaded page to a suspicious domain.

* + 1: URL Length < 54
  + 0: 54 ≤ URL Length ≤ 75
  + -1: URL Length > 75
* **Age of Domain:**

Websites that have an online presence of less than 1 year, can be considered risky.

* + 1: age ≤ 6 months
  + -1: otherwise
* **IP Address in URL:**

Using an IP address in the domain name of the URL is an indicator someone is trying to access the personal information

* + 1: otherwise
  + 0: IP Address exists in URL
* **Classification:**
  + 1: Legitimate
  + 0: Suspicious
  + -1: Phishy

**Resources**

Dataset Information:

<http://fadifayez.com/wp-content/uploads/2017/11/Phishing-detection-based-Associative-Classification-data-mining.pdf>