Malicious and Benign websites dataset

**What the Data is**

The data is a collection of website characteristics. Characteristics are analysed and the values are represented with a type as either malicious or benign.

There are 20 columns of data and a type label for each URL in this dataset.

A sample of the columns the data provides are:

* URL\_LENGTH: the number of characters in the URL
* NUMBER*SPECIAL*CHARACTERS: it is number of special characters identified in the URL, such as, “/”, “%”, “#”, “&”, “. “, “=”
* WHOIS\_STATEPRO: it is a categorical variable, its values are the states we got from the server response.
* DNS*QUERY*TIMES: this is the number of DNS packets generated during the communication between the honeypot and the server
* TYPE: this is a categorical variable, its values represent the type of web page analyzed, specifically, 1 is for malicious websites and 0 is for benign websites.

**Where the Data comes from**

Malicious data was collected from three other sites and verified through VirusTotal (A website owned by Google that analyzes files and URLs for malicious content).

* *machinelearning.inginf.units.it/data-andtools/hidden-fraudulent-urls-dataset*

This dataset includes three kinds of URLs:

Hidden fraudulent URLs,

URLs of legitimate pages belonging to trusted, yet compromised web sites,

URLs of legitimate pages belonging to trusted and uncompromised web sites.

This dataset considers two categories of fraudulent web page:

Web defacements and

Phishing.

Concerning phishing, used the **data provided by Phishtank.** (A free community site where anyone can submit, verify, track and share phishing data) Composed a list of about 7500 valid and online URLs extracted from Phishtank.

Concerning defacements, **data provided by Zone-H** (an archive of defaced websites). Composed a list of about 2500 URLs extracted from Zone-H.

Augmented the lists by adding all URLs of pages reached by crawling the compromised sites and dropped from the lists the following items:

URLs whose domain is an IP address;

URLs whose path is empty or equal to index.html

Concerning URLs of legitimate pages belonging to trusted and uncompromised, we selected a set of 20 web sites extracted from the top 500 web sites ranking provided by Alexa. We excluded from this selection:

web sites providing different content depending on whether the user is authenticated,

social network web sites,

search engines.

For each of these 20 sites, saved all the URLs obtained.

* malwaredomainlist.com

(Malware Domain List (MDL) is a non-commercial community project that lists domains that are known to generate spam, host botnets, create DDoS attacks, and generally contain malware.)

* zeuztacker.abuse.ch

Discontinued open source website that provided malicious websites for blocking.

Benign URLs from (https://github.com/faizann24/Using-machinelearning-to-detect-malicious-URLs.git) verification process also.

(Git page also discontinued)

**What the Data means**

The data explores what characteristics determine a websites type.

The data tracked both application layer and network layer features of labelled URLs.

This provides insights such as:

The number of special characters for Malicious website is higher than benign ones &

the mean length of URL is higher for Malicious website as compared to benign ones.

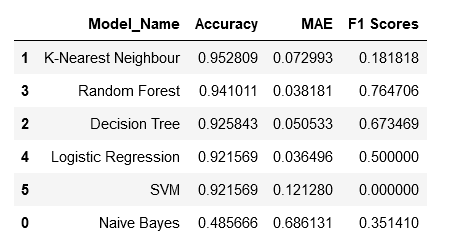
**Predicting**

We have trained the data and used various ML model that predict whether a website is malicious or not.

**Model results**

|  |  |
| --- | --- |
| **Naïve Bayes**  **Accuracy:** 0.4856661045531197 | **K- Nearest Neighbour**    **Accuracy:** 0.7968397291196389 |
| **Decision Tree**    **Accuracy:** 0.9282511210762332 | **Random Forest**    **Accuracy:** 0.9410112359550562 |
| **Logistic Regression**    **Accuracy:** 0.8879551820728291 | **SVM**    **Accuracy:** 0.8375350140056023 |

**Final Metrics**



**Recommended model:** Random Forest