# Springboard Capstone Project Proposal

## Prediction of Phishing Websites

## Introduction

“Phishing” is an instance of social engineering techniques used to deceive users into giving their sensitive information using an illegitimate website that looks and feels exactly like the target organization website. The word phishing is a neologism created as a homophone of fishing due to the similarity of using a bait in an attempt to catch a victim.

Website phishing is considered one of the crucial security challenges for the online community due to the massive numbers of online transactions performed on a daily basis. According to the 2013 Microsoft Computing Safety Index, released in February 2014, the annual worldwide impact of phishing could be as high as US$5 billion.(IANS. news.biharprabha.com. Retrieved February 11, 2014.)

Most phishing detection approaches utilizes Uniform Resource Locator (URL) blacklists or phishing website features combined with machine learning techniques to combat phishing. Despite the existing approaches that utilize URL blacklists, they cannot generalize well with new phishing attacks due to human weakness in verifying blacklists, while the existing feature-based methods suffer high false positive rates and insufficient phishing features. As a result, this leads to an inadequacy in the online transactions.

## Clients

The phishing problem is considered a vital issue especially in e-banking and e-commerce industry by taking the number of online transactions involving payments. Many leading security companies like BitDefender, Symantec, McAfee, VeriSign, IronKey and Internet Identity work to prevent the Phishing.

## Data Set Information:

This dataset has identified different features related to legitimate and phishy websites and collected 2456 different websites. This [Data set](http://archive.ics.uci.edu/ml/datasets/Phishing+Websites) is avaliable at UCI Machine learning repository. Phishing websites were collected from Phishtank data archive (www.phishtank.com),MillerSmiles archive, Google searching operators which are a free community sites where users can submit, verify, track and share phishing data. This dataset has 1148 legitimate websites out of 2456 websites and rest are considered as phishing websites.

### 1. Address Bar based Features

* having\_IP\_Address { -1,1 }
* URL\_Length { 1,0,-1 }
* Shortining\_Service { 1,-1 }
* having\_At\_Symbol { 1,-1 }
* double\_slash\_redirecting { -1,1 }
* Prefix\_Suffix { -1,1 }
* having\_Sub\_Domain { -1,0,1 }
* SSLfinal\_State { -1,1,0 }
* Domain\_registeration\_length { -1,1 }
* Favicon { 1,-1 }
* port { 1,-1 }
* HTTPS\_token { -1,1 }

### 2. Abnormal Based Features

* Request\_URL { 1,-1 }
* URL\_of\_Anchor { -1,0,1 }
* Links\_in\_tags { 1,-1,0 }
* SFH { -1,1,0 }
* Submitting\_to\_email { -1,1 }
* Abnormal\_URL { -1,1 }

### 3. HTML and JavaScript based Features

* Redirect { 0,1 }
* on\_mouseover { 1,-1 }
* RightClick { 1,-1 }
* popUpWidnow { 1,-1 }
* Iframe { 1,-1 }

### 4. Domain based Features

* age\_of\_domain { -1,1 }
* DNSRecord { -1,1 }
* web\_traffic { -1,0,1 }
* Page\_Rank { -1,1 }
* Google\_Index { 1,-1 }
* Links\_pointing\_to\_page { 1,0,-1 }
* Statistical\_report { -1,1 }

### Conclusion

* Result { -1,1 }

### Notations

* -1 <- Phishing
* 0 <- Suspicious
* 1 <- Legitimate

# Approach

For the first step, I will be splitting the data into three parts. One part is for training the model, second part is for the cross validation and the final part for testing the model.

The next step is to exploratory the training data and analyzing the relationships between any of the independent variables with the result. For example in this dataset Since the features was categorized into four groups. I am planning to see if the prediction can be done basing on the categories like how many features in each of the categories need to be positive so that the website can be legitimate.

However there is a easy and direct way which is by considering all the features irrespective of the catogaries and try to build a model considering all the features.

# Deliverables

The deliverables for this project will be uploaded onto github, which includes the code used for Data Wrangling, Exploring and analyses. In the end there will be a detail report which includes analysis and final conclusion.