Case Study Proposal, IE7275

Group 04 – Ronit Mankad, Shashank Nukala

Background

Phishing is a form of cyber-attack that utilizes counterfeit websites to steal sensitive user information such as account login credentials, credit card numbers, etc. Throughout the world, phishing attacks continue to evolve and gain momentum. In June 2018, the Anti-Phishing Working Group (APWG) reported as many as 51,401 unique phishing websites.

Phishing is an example of social engineering techniques being used to deceive users. Users are often lured by communications purporting to be from trusted parties such as social web sites, auction sites, banks, online payment processors or IT administrators. Attempts to deal with phishing incidents include legislation, user training, public awareness, and technical security measures (the latter being due to phishing attacks frequently exploiting weaknesses in current web security).

Cyber Security researchers have made great strides in developing tools and methods to detect and report phishing websites. The advent of machine learning methods has also helped in developing applications which can automatically detect phishing attacks.

Data Collection

Data Source:

The data was obtained from the following source -

[1] Tan, Choon Lin (2018), "Phishing Dataset for Machine Learning: Feature Evaluation", Mendeley Data, v1

Data Description:

This dataset contains 48 features extracted from 5000 phishing webpages and 5000 legitimate webpages, which were downloaded from January to May 2015 and from May to June 2017. An improved feature extraction technique is employed by leveraging the browser automation framework (i.e, Selenium WebDriver), which is more precise and robust compared to parsing approach based on regular expressions.

Phishing webpage source: PhishTank, OpenPhishLegitimate webpage source: Alexa, Common Crawl

The Problem

Identifying phishing websites across different platforms still proves to be a major challenge in the industry. Websites can be classified by monitoring a variety of different indicators. Example: does the website uses https or not, does the website uses an external favicon etc. As the number of indicators increase, they introduce more complexity to the classification process.

Possible Solution

We propose a solution which uses machine learning techniques like KNN, CART, Naive Bayes, Logistic regression, Neural Nets, Latent Discriminant Analysis and Support Vector Machine to model and classify the data. A well trained and generalized model will be able to classify websites with a reasonable accuracy. Furthermore, we will also implement dimension reduction techniques like Principal Component Analysis to streamline the dataset. The models will be tested for their accuracy and robustness using evaluation metrics like Lift Chart and ROC curve.

References

[1] Tan, Choon Lin (2018), "Phishing Dataset for Machine Learning: Feature Evaluation", Mendeley Data, v1

[2] Wikipedia contributors. (2020, May 21). Phishing. In Wikipedia, The Free Encyclopedia. Retrieved 15:42, May 24, 2020