Phishing Website Detection by Machine Learning Techniques

Team Members
Aman Mittal (am11982)
Kara Vora (kv2154)
Siddharth Shah (ss16130)
Parth Metha (pjm9767)

Objective:

A phishing website is a common social engineering method that mimics trustful uniform resource locators (URLs) and web pages. The objective of this project is to train machine learning models and deep neural nets on the dataset created to predict phishing websites. Both phishing and benign URLs of websites are gathered to form a dataset and from them required URL and website content-based features are extracted. The performance level of each model is measured and compared.

Datasets:

The set of phishing URLs is collected from an open-source service called PhishTank. This service provides a set of phishing URLs in multiple formats like CSV, JSON, etc. that get updated hourly. To download the data: https://www.phishtank.com/developer_info.php. From this dataset, 5000 random phishing URLs are collected to train the ML models.

The legitimate URLs are obtained from the open datasets of the University of New Brunswick, https://www.unb.ca/cic/datasets/url-2016.html. This dataset has a collection of benign, spam, phishing, malware & defacement URLs. Out of all these types, the benign URL dataset is considered for this project. From this dataset, 5000 random legitimate URLs are collected to train the ML models.

Feature Engineering:

URL-Based Features:

-> Length of URL: Longer URLs may be suspicious.
-> URL Shortening: Use of URL shortening services.
-> Presence of IP Address: URLs containing IP addresses could indicate phishing.
-> Number of Subdomains: Excessive use of subdomains might be suspicious.
-> Presence of HTTPS: Whether the URL uses HTTPS or not.
-> Tokenization: Breaking the URL into meaningful tokens (words, numbers, characters).

Domain-Based Features:

-> Domain Registration Length: Short-term domain registration can be suspicious.

-> Domain Age: Newer domains might be more suspect.
-> DNS Record: The absence of DNS records could be a red flag.
-> WHOIS Information: Privacy-protected WHOIS information could be indicative of phishing.

HTML and JavaScript-Based Features:

-> Use of iFrames: Phishing sites often use iFrames to display content from legitimate sites.
-> Presence of Forms: The presence and number of input forms, especially those that transmit data.
-> JavaScript Obfuscation: The use of obfuscated JavaScript code can be a sign of malicious intent.

Model Selection:

Before starting the ML model training, the data is split into 80-20 i.e., 8000 training samples & 2000 testing samples. From the dataset, it is clear that this is a supervised machine-learning task. Two major types of supervised machine learning problems are classification and regression.

This data set comes under a classification problem, as the input URL is classified as phishing (1) or legitimate (0). The supervised machine learning models (classification) considered to train the dataset in this project are:

Decision Tree
Random Forest
Multilayer Perceptrons
XGBoost
Autoencoder Neural Network
Support Vector Machines