Detection of Phishing Websites using feature extraction and machine learning techniques.

Btech CSE 2024

By

Shubham Mandal

Introduction

On the era of internet cybercrimes has been incresed by many folds making it more vulnerable to data stealing and leaking of credentials for the solution of which i made an Supervised machine learning model using very well known Algorithmns comparing the common malpractices defined as features and result the output as in the terms of Accuracy and Performance

Literature Review

- For this project in particular i had studied the recent records of cyber crime and reviewed few articles regarding the same from the site of university of burnswick and the other reports from NullByte and phishtank.

Methodology

Collected dataset containing phishing and legitimate websites from the open source platforms.

- Wrote a code to extract the required features from the URL database.
- Divide the dataset into training and testing sets.
- Run selected machine learning and deep neural network algorithms like SVM, Random Forest,Autoencoder on the dataset.
- Wrote a code for displaying the evaluation result considering accuracy metrics.
- Compare the obtained results for trained models and specify which is better.

Data Collection

- Legitimate URLs are collected from the dataset provided by University of New Brunswick, https://www.unb.ca/cic/datasets/url-2016.html.
- From the collection, 5000 URLs are being randomly picked.
- Phishing URLs are collected from opensource service called PhishTank . This service provide a set of phishing URLs in multiple formats like csv, json etc that gets updated hourly.
- Form the obtained collection, 5000 URLs are being randomly picked.

Feature Selection

The following category of features are selected:

Address Bar based Features considered are:

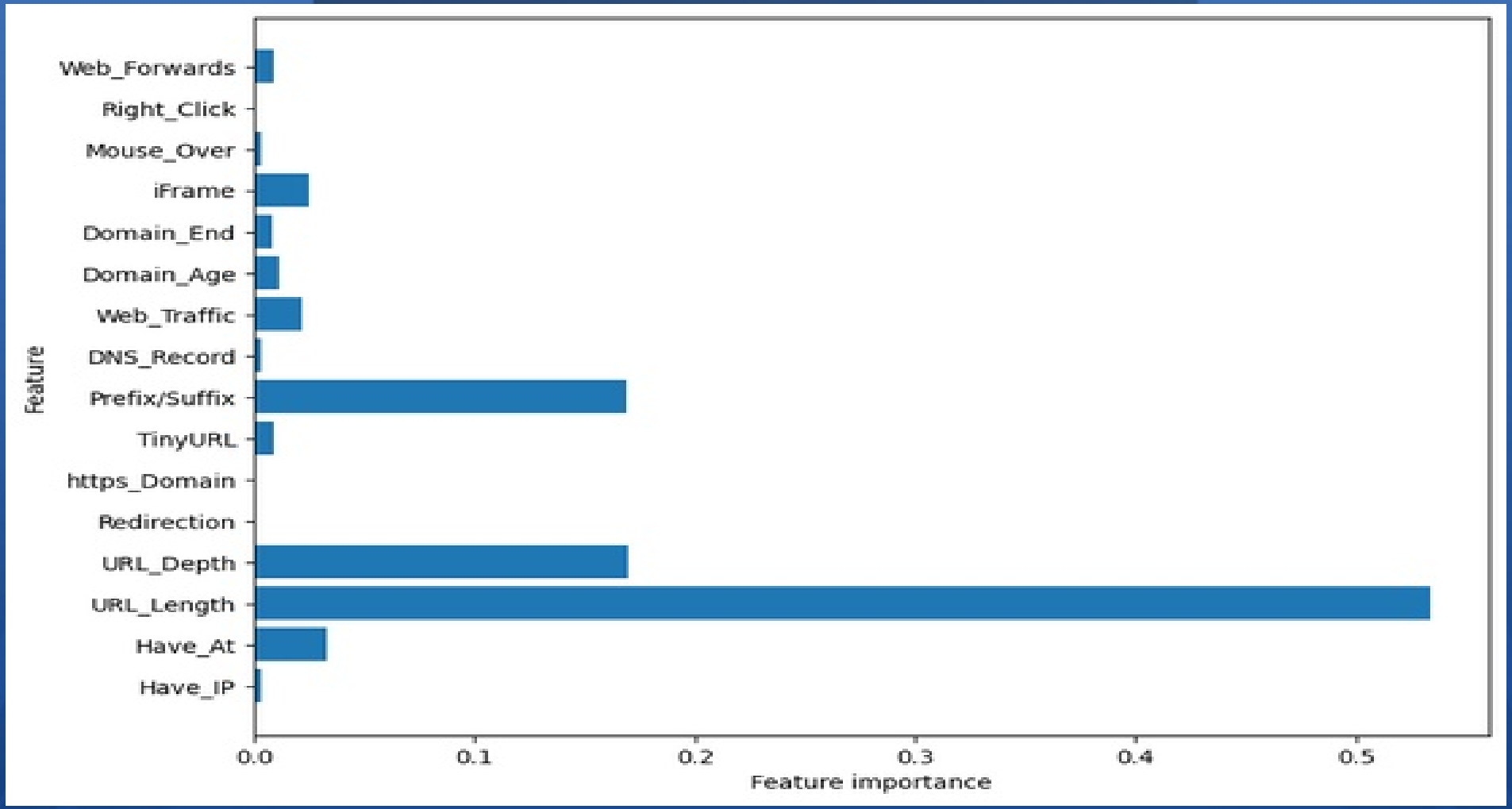
- Domian of URL
- IP Address in URL
- '@' Symbol in URL
- Length of URL
- Depth of URL
- Redirection '//' in URL
- -'http/https' in Domain name
- Using URL Shortening Service
- -Prefix or Suffix "-" in Domain

Domain based Features considered are:

- DNS Record
- Age of Domain
- Website Traffic
- End Period of Domain

HTML and JavaScript based Features considered are:

- Iframe Redirection
- Disabling Right Click
- Status Bar Customization
- Website Forwarding


Web_Forwards
Right_Click
Mouse_Over
iFrame
Domain_End
Domain_Age
Web_Traffic
DNS_Record
Prefix/Suffix
TinyURL
https_Domain
Redirection
URL_Depth
URL_Length
Have_At
Have_IP
Feature
0.0
0.1
0.2
0.3
0.4
0.5
Feature importance

Machine Learning Models

This is a supervised machine learning task. There are two major types of supervised machine learning problems, called classification and regression.

This data set comes under classification problem, as the input URL is classified as Phishing or Legitimate The machine learning models (classification) considered to train the dataset in this notebook are:

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

Model Evalution

	ML Model	Train Accuracy	Test Accuracy
0	Decision Tree	0.816	0.804
1	Random Forest	0.821	0.809
2	Multilayer Perceptrons	0.861	0.845
3	XGBoost	0.867	0.858
4	AutoEncoder	0.753	0.756
5	SVM	0.804	0.793

Conclusion

In closing, this study demonstrates that combining machine learning with careful feature extraction offers a powerful tool to fight phishing attacks. By examining website characteristics like web addresses, content, and underlying code, machine learning models can be trained to identify fake websites. This approach is adaptable, allowing for the inclusion of new features and training on ever-changing phishing tactics. As online security remains a top priority, robust phishing detection systems powered by machine learning will be key to protecting users and creating a safer online environment.