University of New Haven

PHISHING WEBSITE DETECTION by MACHINE LEARNING TECHNIQUES

SHREYA GOPAL SUNDARI

INTRODUCTION

- Phishing is the most commonly used social engineering and cyber attack.
- Through such attacks, the phisher targets naïve online users by tricking them into revealing confidential information, with the purpose of using it fraudulently.
- In order to avoid getting phished,
    - users should have awareness of phishing websites.
    - have a blacklist of phishing websites which requires the knowledge of website being detected as phishing.
    - detect them in their early appearance, using machine learning and deep neural network algorithms.
- Of the above three, the machine learning based method is proven to be most effective than the other methods.
- Even then, online users are still being trapped into revealing sensitive information in phishing websites.

OBJECTIVES

A phishing website is a common social engineering method that mimics trustful uniform resource locators (URLs) and webpages. 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 measures and compared.

APPROACH

Below mentioned are the steps involved in the completion of this project:

- Collect dataset containing phishing and legitimate websites from the open source platforms.
- Write a code to extract the required features from the URL database.
- Analyze and preprocess the dataset by using EDA techniques.
- 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.
- Write 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 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 randomly picked.

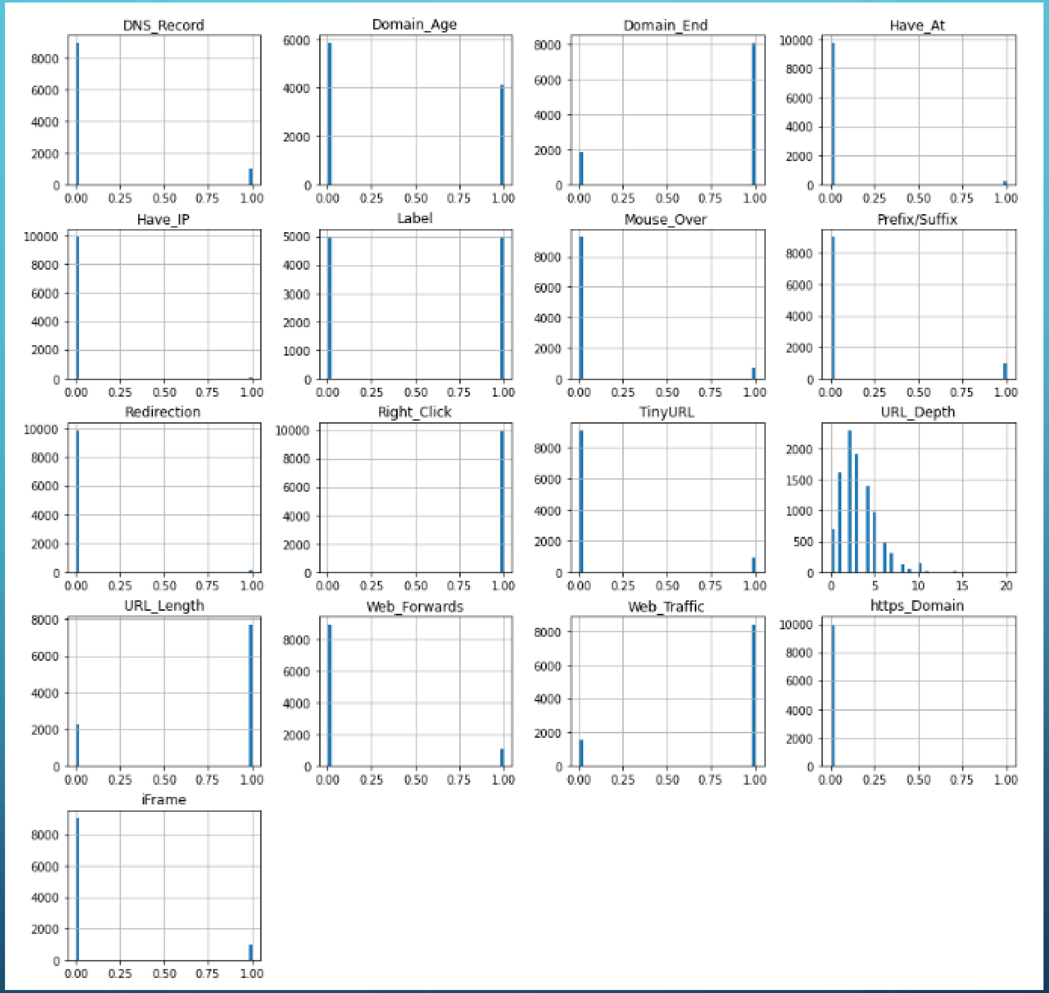
FEATURE SELECTION

- The following category of features are selected:
  - Address Bar based Features
  - Domain based Features
  - HTML & Javascript based Feature
- 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

FEATURE SELECTION (CONT.)

- Domain based Features considered are:
  - DNS Record
  - Website Traffic
  - Age of Domain
  - End Period of Domain
- HTML and JavaScript based Features considered are:
  - Iframe Redirection
  - Status Bar Customization
  - Disabling Right Click
  - Website Forwarding
- All together 17 features are extracted from the dataset.

FEATURES DISTRIBUTION

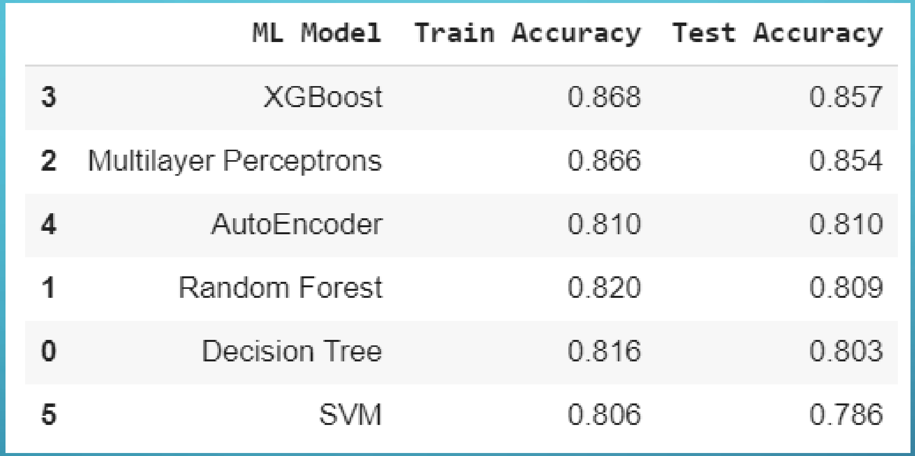


MACHINE LEARNING MODELS

- 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 (1) or legitimate (0). 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 EVALUATION

- The models are evaluated, and the considered metric is accuracy.
- Below Figure shows the training and test dataset accuracy by the respective models:



	ML Model	Train Accuracy	Test Accuracy
3	XGBoost	0.868	0.857
2	Multilayer Perceptrons	0.866	0.854
4	AutoEncoder	0.810	0.810
1	Random Forest	0.820	0.809
0	Decision Tree	0.816	0.803
5	SVM	0.806	0.786

- For the above it is clear that the XGBoost model gives better performance. The model is saved for further usage.

University of New Haven

NEXT STEPS

- Working on this project is very knowledgeable and worth the effort.
- Through this project, one can know a lot about the phishing websites and how they are differentiated from legitimate ones.
- This project can be taken further by creating a browser extensions of developing a GUI.
- These should classify the inputted URL to legitimate or phishing with the use of the saved model.

University of New Haven

Thank You.....