Phishing website Detection by Machine Learning

Ariga Tejasimha 12/05/2021

Content

- Introduction
- Objective
- Approach
- Data Collection
- Feature Collection
- ML Models
- Model Evaluation

Introduction

- Phishing the most Commonly used Social Engineering and Cyber Attack
- With this attacks they target the online user by tricking them to reveal their sensitive information with the purpose of fraudulent
- To Avoid gutted Phished user must have good knowledge of these websites, Must have blacklisted these kind of websites or using machine learning techniques to detect them earlier

Objective

A Phishing website is nothing but mimics the trustful URL and Webpages. Our Objective of the project is to train the ML and Deep Neural Nets on the Dataset created to predict phishing websites. We gonna mix both Phishing and benign URLs are formed to a dataset. From which we gonna extract the required URL and website content based featured are extracted. The performance of each Machine Learning Model is compared

Approach

- Collecting the Dataset of both Phishing and Legitimate websites
- Creating a required Dataset of required URLs
- Preprocessing the Dataset that we acquired
- Dividing the Dataset into training and testing sets
- Building the ML Models Like Decision Tree, Random Forest logistic regression E.t.c,
- Evaluation of the Models using the Metric Accuracy
- Comparing the models

DataCollection

- We Randomly pick Legitimate (University of New Brunswick)and Phishing (Phishing Tank)URLs 5000 from respective Open Source Platforms.
- As they are in their respective formats we will process them into List format first and then Dataframes and then finally into CSV.
- We will obtain the Csv Dataset which we will use to Build the ML and Deep Neural Net Models on it.
- Dataset total has 1000 entries with 14 columns of Features which is explained in the following the slides

Feature Selection

- We are extracting using only two categories they are
  - Address Bar Based Features
  - Domain Based Features
- Address Bar Based Features are

1. Domain 2. Redirection ‘//‘ in URL 3. IP address

4. ‘http/https’ in domain name 5. ‘@‘ Symbol in URL.

6. URL shortening Service. 7. lenght of URL.

8.Depth of URL

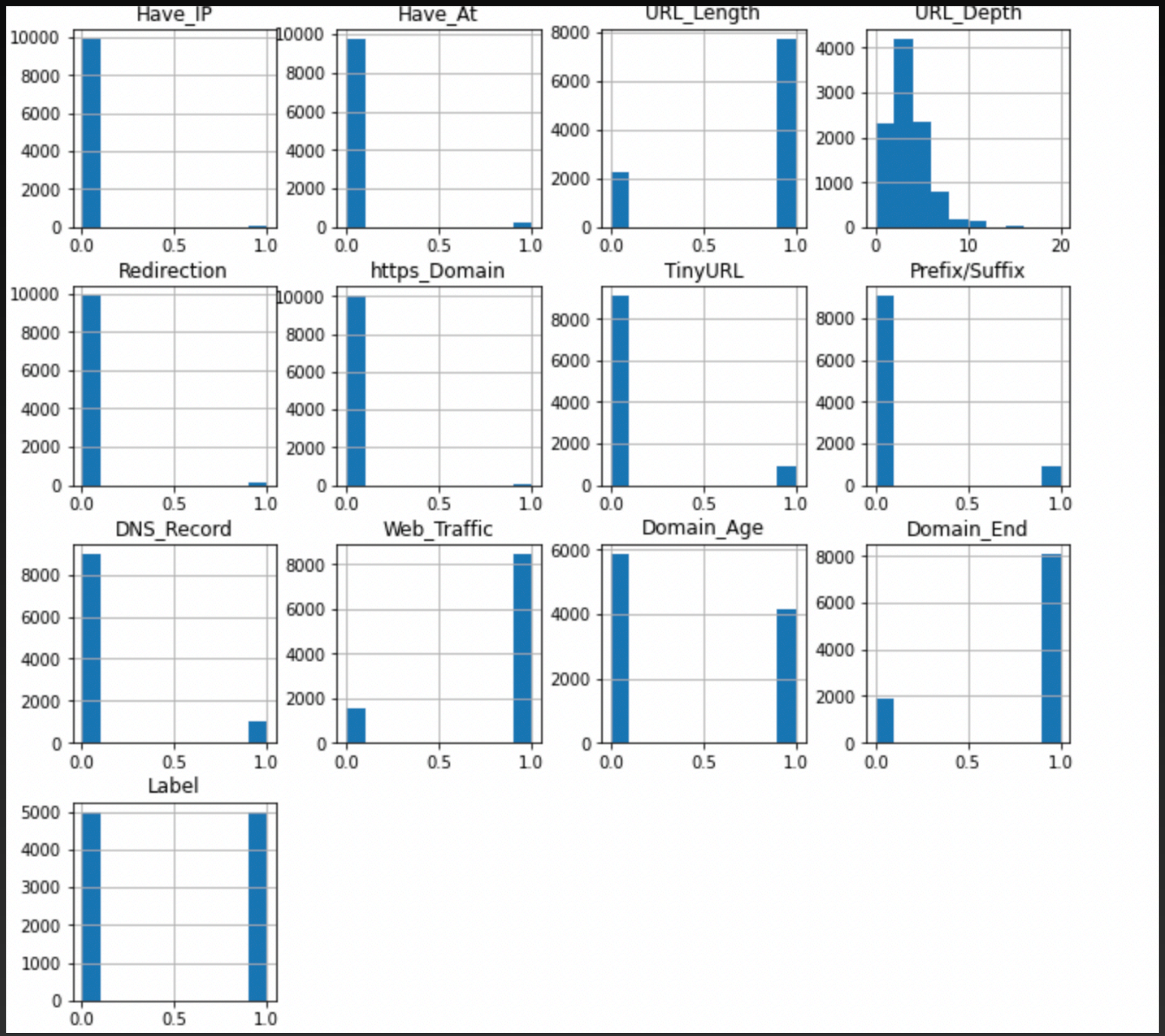
9.prefix or Suffix ‘-‘ in Domain

Feature Selection

- Domain Based Feature extraction
    - DNS Record
    - Age of Domain
    - Website Traffic
    - End period of Domain

All together they are 13 Features extracted with 14 column being the Label with ‘0’ being the legitimate and ‘1’ being the phishing

Feature Distribution



Machine Learning

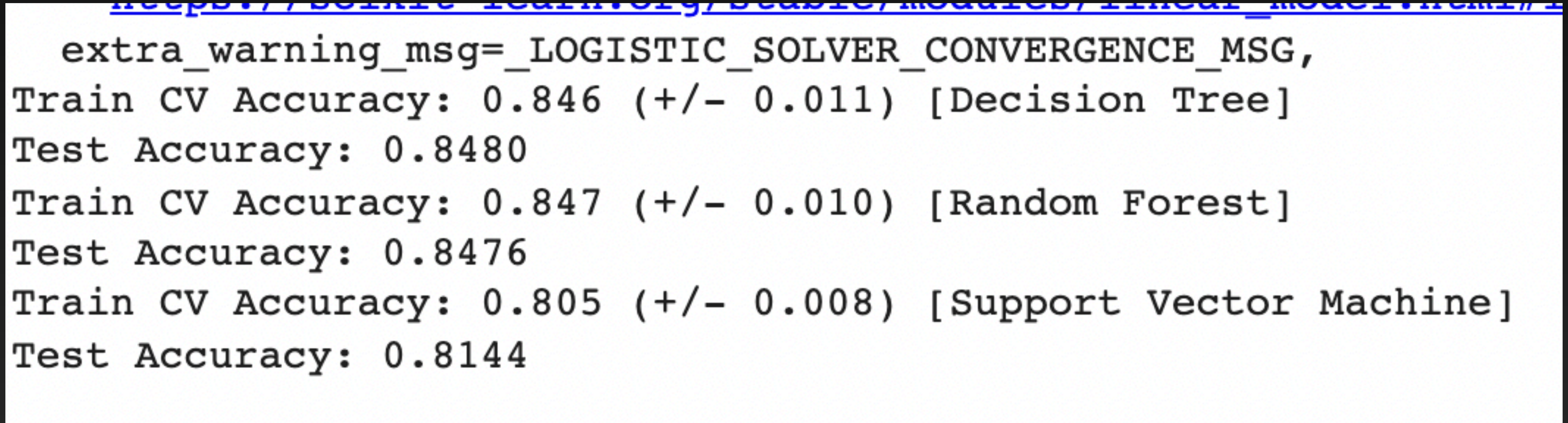
After getting the require dataset we have build some of the machine learning Models.

From the dataset we have done the URL that is phishing (1) and legitimate (0). We have done the classification. We have build some of the models like

- Logistic Regression
- Decision Tree
- Random Forest
- SVM

Model Evaluation

- The Models are build and Metric we used is Accuracy
- I have done 5-cross folding to Get the best out of the models



```
    https://scikit-learn.org/stable/modules/linear_model.html#l
  extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG,
Train CV Accuracy: 0.846 (+/- 0.011) [Decision Tree]
Test Accuracy: 0.8480
Train CV Accuracy: 0.847 (+/- 0.010) [Random Forest]
Test Accuracy: 0.8476
Train CV Accuracy: 0.805 (+/- 0.008) [Support Vector Machine]
Test Accuracy: 0.8144
```

- From this we can say that Decision Tree has best output of it

Summary

- Working on this project and taking this course increase my scope and perpective of seeing the problem and world has changed. Coming across the technique or word Zero Day attack. Completely changed my thought process
- After going through lot of research papers as a part of circullam
- The Further step is so simple and many of the search engines or wifi provider are using this solutions. They are allowing the access of such kind of content before we can able to access it.