PROJECT REPORT - PHISHING URL DETECTION

BY - SMAI234

Team Members:

1. Rajneesh Singhatiya
2. Souparna Das
3. Archit Kumar
4. Aman Agarwal

Assigned TA : Satyam Mittal

Literature Survey :

Paper reference : https://arxiv.org/pdf/1701.07179.pdf

Data Extraction :

https://www.kaggle.com/xwolf12/malicious-and-benign-websites

Data Analysis :

Data present at above lin has following columns :

URL
URL_LENGTH
NUMBER_SPECIAL_CHARACTERS
CHARSET
SERVER
CONTENT_LENGTH
WHOIS_COUNTRY
WHOIS_STATEPRO
WHOIS_REGDATE
WHOIS_UPDATED_DATE
TCP_CONVERSATION_EXCHANGE
DIST_REMOTE_TCP_PORT
REMOTE_IPS
APP_BYTES
SOURCE_APP_PACKETS
REMOTE_APP_PACKETS
SOURCE_APP_BYTES
REMOTE_APP_BYTES
APP_PACKETS
DNS_QUERY_TIMES
Type

Problem Categorisation :

Malicious URL Detection using Machine Learning : Malicious URL, a.k.a. malicious website, is a common and serious threat to cybersecurity. Malicious URLs host unsolicited content (spam, phishing, drive-by exploits, etc.) and lure unsuspecting users to become victims of scams (monetary loss, theft of private information, and malware installation), and cause losses of billions of dollars every year. It is imperative to detect and act on such threats in a timely manner.

Success Metric :

Accuracy, AUROC, precision, recall etc.

Feature Extraction :

Types of features :

1. URL-Based Features
2. Domain-Based Features
3. Page-Based Features
4. Content-Based Features

Model Selection :

Different models will be tried and tested after that we will finalize the model to be used.

Validation and Testing :

Has not been decided yet.

Github Repo :

https://github.com/agarwal29796/fraud_page_detection