Phishing Website Detection Using Machine Learning

Prof. Vaishnavi Ganesh
Department of Computer Science and Engineering
Priyadarshini College of Engineering
Nagpur, India

Anjali Sulakhiya
Department of Computer Science and Engineering
Priyadarshini College of Engineering
Nagpur, India

Jay Dhurat
Department of Computer Science and Engineering
Priyadarshini College of Engineering
Nagpur, India

Kaustubh Shastrakar
Department of Computer Science and Engineering
Priyadarshini College of Engineering
Nagpur, India

Shrutika Satange
Department of Computer Science and Engineering
Priyadarshini College of Engineering
Nagpur, India

Nikhil Miralwar
Department of Computer Science and Engineering
Priyadarshini College of Engineering
Nagpur, India

---***---

Abstract* - *Phishing websites, which pose as reputable websites and mislead unwary visitors into disclosing sensitive information, are one of the main sources of security breaches in the modern digital era. The easiest method of obtaining sensitive information from unwitting people is through a phishing attack. The goal of phishers is to get crucial data, such as login, password, and bank account information. By taking advantage of the user's vulnerability, a hacker may be able to obtain information such as bank account numbers, passwords to social media accounts, firm income reports, and the details of online transactions, to name a few. This research seeks to identify phishing URLs and identify the most effective machine learning approach based on precision, false-positive rate, and false-negative rate.

I. INTRODUCTION

Phishing is the practise of tricking an individual through an electronic connection in order to get sensitive data like usernames, passwords, and credit card numbers. Customers are frequently encouraged to input personal information on a fake website that looks and feels precisely like the genuine one through email spoofing or instant messaging, which is how it's generally done. One of the most harmful and hazardous illegal activities that is expanding in online. Users who utilise the internet to obtain the services it provides have been quickly falling victim to phishing assaults over the past several years on purpose.

In order to collect sensitive information, the crooks first make unofficial copies of legitimate websites and emails, typically from banking institutions or other businesses that deal with financial information. The words and logos of an authentic firm will be used to construct the email. One of the factors contributing to the Internet's fast expansion as a communication medium is the nature of website construction, which also makes it possible to misuse the trademarks, trade names, and other corporate identifiers that customers have come to rely on as procedures for identification.

The "spoofed" emails are then distributed to as many individuals as possible in an effort to deceive them. Customers are routed to a fake website that seems to be from the real company when they receive these emails or click a link in them.

Internet users are at risk from a number of cyber dangers, such as identity theft, theft of personal information, and financial losses. As a result, internet usage at home and at work can be dubious. Users should be able to recognise and protect against privacy leaks using efficient analytical tools in order to reduce security risks. An information security management system based on artificial intelligence should be used to construct efficient systems that can enhance self-intervention at the moment of an attack.

II. Literature Survey

Phishing is a technique used to steal data, money, or personal information using a false website. The greatest method for preventing contact with the phishing website is to identify dangerous URLs in real-time. Identifying phishing websites depends on their domains. They often have something to do with URLs (low-level and upper-level domains, paths, and queries) that need to be registered.

Utilizing distinguishing features taken from the words that make up a URL and query data from several search engines, like Google and Yahoo, it is possible to evaluate recently acquired state of intra-URL relationships. The machine-learning based classification for the detection of phishing URLs from a real dataset is further influenced by these attributes.

This research uses phish-STORM to focus on real-time URL phishing versus phishing material. In order to distinguish between phishing and non-phishing URLs, a few relationships between the registration domain and the remainder of the URL are taken into consideration for this. Certain common blacklisted urls are used to identify phishing websites, although this method is ineffective because phishing websites only exist for a brief period of time. The practise is known as phishing. It is the practise of misleading a company's clients to communicate with their sensitive information in an unethical manner. It is also possible to describe it as the deliberate use of harsh tools like spam to automatically target the victims and collect their private information.

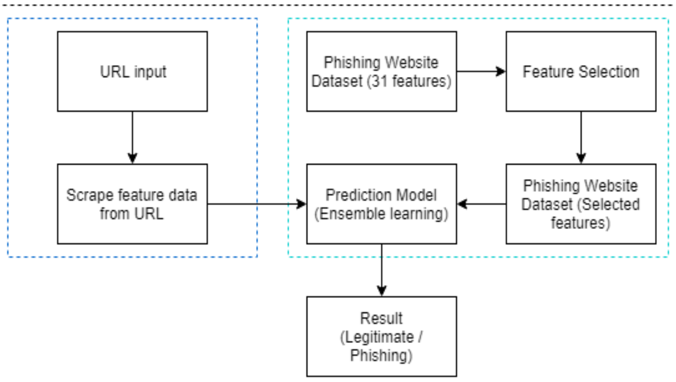
There is more communication available for the delivery of malicious messages since many of the SMTP failures are exploitation channels for phishing websites. A brand-new feature extraction method for categorization that uses heuristics was proposed. In this, they have categorised extracted characteristics into many categories, such as features that obfuscate URLs and features that are dependent on hyperlinks. Additionally, the suggested approach provides 92.5% accuracy. Additionally, the number, quality, and feature extraction of the training data are the only factors that affect this model.

III. Methodology

Our project was created utilising a website that serves as a software for all users. It will be possible to tell whether a website is real or phishing by using this engaging and responsive website.

React was used to create this website. React is a front-end JavaScript toolkit that is free and open-source for creating user interfaces based on UI components. It should be emphasised that the website is intended for all users, thus it must be simple to use and no user should encounter any difficulties.

The website provides details about the services we offer. It also includes information about unethical behaviours occurring in the technology world of today. The website is designed with the intention of educating users about the malpractices taking place in today's society as well as enabling them to distinguish between authentic and fake websites. They can avoid someone attempting to utilise their personal information, such as their email address, password, debit card number, credit card number, CVV number, bank account number, and so on.


URL input
Scrape feature data from URL
Phishing Website Dataset (31 features)
Feature Selection
Prediction Model (Ensemble learning)
Phishing Website Dataset (Selected features)
Result (Legitimate / Phishing)

1) *Dataset Import*

Import a dataset from Kaggle.com that contains both genuine and phishing URLs, designated as "0" for trustworthy websites and "1" for malicious websites.

2) *Data preprocessing*

includes purging, instance picking, feature extraction, normalisation, transformation, etc. The entire training dataset is the end result of data preparation. Data pretreatment may influence how the final processing's results are interpreted. Data cleaning might involve filling in the gaps in the data, reducing noise, identifying and eliminating outliers, and addressing incompatibilities. A technique for adding precise databases or data sets is called data integration. Data transformation is the process of gathering and normalising data in order to measure a certain set of data. By doing data reduction, we may provide a very brief summary of the dataset that nonetheless contributes to the same analytical result.

3) *Trained ML Model*

Used Google Colab to train the model with features such as:

- URL redirection: If the URL path contains "//," the feature is set to 1; otherwise, it is set to 0. The visitor will be moved to another website if the URL path contains the symbol "/".
- Length of Host name: The average length of benign URLs is 25, and if the URL is longer than 25, the feature is set to 1; otherwise, it is set to 0.
- URL Shorten Services "TinyURL": With the use of the TinyURL service, phishers may disguise lengthy phishing URLs as small ones. User traffic is being diverted to fraudulent websites. If the URL is shortened using a service like bit.ly, then the feature is set to 1, otherwise it is set to 0.

- Existence of @ symbol in URL: If the URL contains the @ sign, the feature is set to I; otherwise, it is set to 0. When phishers add a specific @ sign to a URL, the browser ignores everything before the "@" symbol and frequently skips to the true address after it.

- Existence of IP address in URL: The feature is set to 1 if the URL contains the IP address; otherwise, it is set to 0. The majority of trustworthy websites never use an IP address as the URL to download a webpage. The use of an IP address in a URL suggests that the attacker is attempting to steal sensitive data.

- Information submission to Email: Using the "mail ()" or "mailto:" methods, the phisher can send the user's data to his own email. If the URL contains such functions, the feature is set to 1; otherwise, it is set to 0.

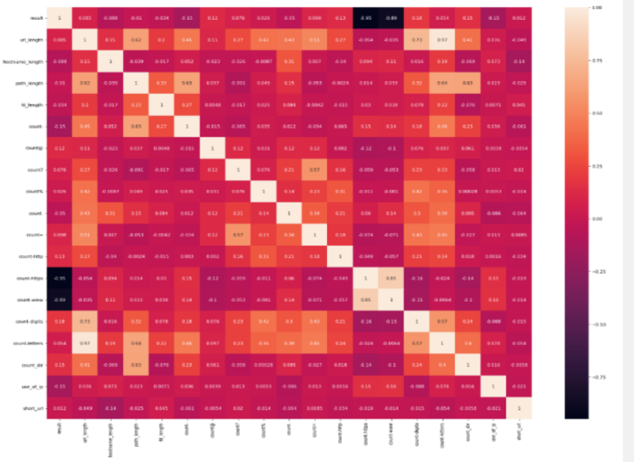
- Number of slash in URL: The average number of slashes in benign URLs is 5. If that number is higher, the feature is set to 1; otherwise, it is set to 0.

- URL of Anchor: You have obtained this functionality by crawling the URL and its source code. The a> element specifies the URL of the anchor. The feature is set to 1 if the a> tag has a maximum number of hyperlinks from another site; otherwise, it is set to 0.

MLP, or a multilayer perceptron, was employed. Another name for multi-layer perception is MLP. It is made up of thick, fully linked layers that may change any input dimension into the required dimension. A neural network with numerous layers is referred to as a multi-layer perception. In order to build a neural network, we join neurons so that some of their outputs are also their inputs.

4) *Exploratory Data Analysis*

Exploratory Data Analysis (EDA) is a method of data analysis that offers several methods and is largely diagrammatic, as seen below. It enhances the perception of a data collection, reveals the underlying structure, extracts key parameters, finds outliers and abnormalities, and tests the Heatmap's hidden audacity.



5) *Develop API and host it on render.com*

API stands for Application Programming Interface, which is a set of definitions and protocols for building and integrating application software.

APIs are mechanisms that enable two software components to communicate with each other using a set of definitions and protocols.

The source code for our API is at https://github.com/prof-moriarty/fishyapi and it is hosted at render.com.

Render is a unified cloud to build and run all your apps and websites with free TLS certificates, a global CDN, DDoS protection, private networks, and auto deploys from Git. Render deploys the API from github and hosts it on its servers.

6) *Develop a frontend with React and host it on Github Pages*

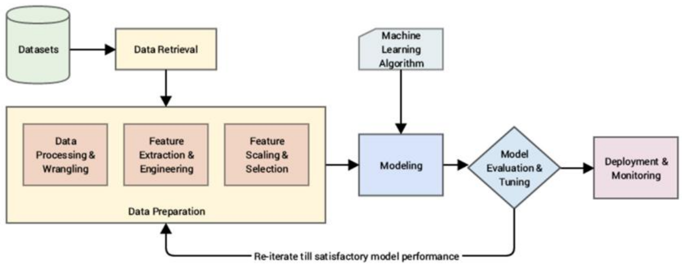
The frontend is made in React. React is a free and open-source front-end JavaScript library for building user interfaces based on UI components.

The frontend for this project is hosted on Github Pages at https://prof- moriarty.github.io/fishy0/

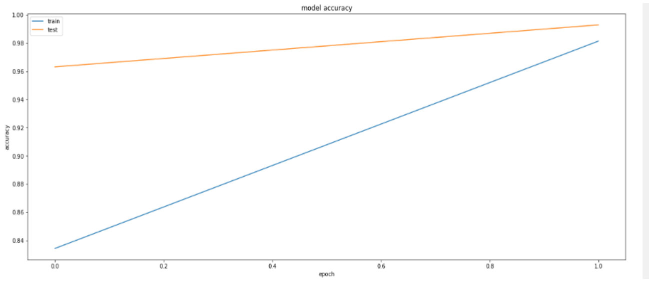
7) *Working*

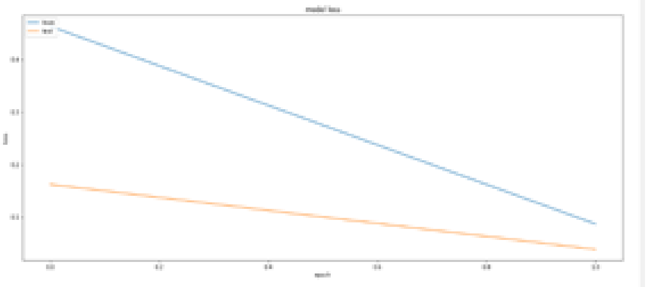
A URL is entered into the search field, and the Scan button is then clicked. The URL is then forwarded to the render.com API, where it is scanned and examined using the model created before.

The input URL will be scanned and separated into real and bogus URLs using a model trained with Multilayer Perceptron. After being scanned, the input URL is given a likelihood score (in the form of %). The cutoff for this likelihood score is 70%. The URL is very likely to take you to a phishing attempt if the score is higher than 70%. The URL is more likely to be secure if the score is under 70%.


Datasets
Data Retrieval
Machine Learning Algorithm
Data Processing & Wrangling
Feature Extraction & Engineering
Feature Scaling & Selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deployment & Monitoring
Re-iterate till satisfactory model performance

8) *Comparison between different algorithm for accuracy*





IV. Conclusion

Phishing is a type of criminal behaviour that uses social engineering methods in the computing industry. The main goal of early phishing attempts was to get access to the victim's AOL accounts, or sporadically to steal credit card information for fraudulent use. The majority of phishing techniques include some kind of technical deception plan to make a link in an email seem to come from the fake company. Significant security issues still exist in reducing the number of unprotected PCs that feed botnets, combating the rise in spam email, stopping organised crime, and warning Internet users about the dangers of social engineering. The objective of this study's future work is to create an unsupervised deep learning technique that can extract knowledge from URLs. The research can also be expanded in order to get results for a bigger network while maintaining an individual's right to privacy.

According to our research, web phishing may be identified by uztilising a classifier and a multilayer perceptron machine learning system. Our study demonstrates that classifiers perform better when more features are used as training data and when the most important characteristics are used as training data. Currently, we have classifiers that detect phishing websites with high accuracy. The provocation in this area will be that scammer will continue to improve the URLs and design of phishing websites so that they resemble legitimate websites. It is thus vital to enhance current features and add new ones for phishing detection.

V. References

[1] Joby James, Sandhya L., Ciza Thomas; "Detection of phishing URLS using machine learning techniques"; International Conference on Control Communication and Computing (ICCC); 2013;

DOI:10.1109/ICCC.2013.6731669

[2] M Selvakumari, M Sowjanya, Sneha Das, S Padmavathi; "Phishing website detection using machine learning and deep learning techniques"; Journal of Physics Conference Series: 2021;

DOI:10.1088/1742-6596/1916/1/012169

[3] Rishikesh Mahajan, Irfan Siddavatam,"Phishing Website Detection using Machine Learning Algorithms; International Journal of Computer Applications"; 2018;

DOI:10.5120/ijca2018918026

[4] Arun Kulkarni, Leonard L. Brown; "Phishing Websites Detection using Machine Learning", International Journal of Advanced Computer Science and Applications (IJACSA), Volume 10, 2019;

(DOI) 10.14569/IJACSA 2019.0100702

[5] Arathi Krishna V, Anusree A, Blessy Jose, Karthika Anilkumar, Ojus Thomas Lee, "Phishing Detection using Machine Leaming based URL Analysis: A Survey, National Conference on Novel & Challenging Issues and Recent Innovations in Engineering and Information Sciences (NCREIS); 2021, DOI: 10.17577/1JERTCONV9IS13033

[6] J. Kumar, A. Santhanavijayan, B.Janet, B. Rajendran and B. S. Bindhumadhava, "Phishing Website Classification and Detection Using Machine Learning," 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1–6, 10.1109/ICCCI48352.2020.9104161.

[7] Hassan Y.A. and Abdelfettah B, "Using case- based reasoning for phishing detection", Procedia Computer Science, vol. 109, 2017, pp. 281–288.

[8] Rao RS, Pais AR. Jail-Phish: An improved search engine based phishing detection system. Computers & Security. 2019 Jun 1;83:246–67.