Project Report: Phishing Website Detection

SAIWARA MAHMUD TUHEE
ID: 20101465
CSE474
Section: 1

May 2023

Introduction

Phishing websites are fraudulent websites designed to trick unsuspecting users into giving up their personal or sensitive information, such as usernames, passwords, and credit card details. Phishing websites can be created by attackers to mimic legitimate websites such as online banking, social media, or e-commerce platforms. These websites often use social engineering tactics to gain the user's trust and persuade them to provide their information.

The consequences of falling victim to phishing websites can be severe. Attackers can use the stolen information to gain unauthorized access to user accounts, make fraudulent purchases, or even steal the victim's identity. To protect against phishing websites, it's essential to understand how they work and how to identify them.

Machine learning can be a powerful tool in detecting phishing websites. The project available at the Kaggle platform, conducted by a user named "buneshathankar25," provides a dataset of over 11,000 website URLs. Each sample includes 30 website parameters and a class label indicating whether it's a phishing website or not. The dataset can be used to train machine learning models for phishing website detection.

By leveraging machine learning algorithms and analyzing the website's features, it's possible to identify patterns that are common to phishing websites. These algorithms can then be used to scan new URLs and classify them as phishing or legitimate. With the increasing sophistication of phishing attacks, machine learning can play a vital role in identifying these malicious websites and protecting users from falling victim to these attacks.

Methodology

The methodology for the project on detecting phishing URLs using machine learning involves the following steps:

Data Collection: The first step involves collecting a dataset of URLs, each labeled as a phishing or legitimate URL. The dataset is available in CSV format and contains 12,000 rows and 31 columns. The dataset provides features such as URL length, domain registration length, and SSL certificate validity that can be used as inputs for the model building process.

Data Preparation: In this step, the dataset is preprocessed and cleaned to remove any missing or irrelevant data. Feature engineering techniques are applied to extract meaningful features that can be used in the machine learning model. The dataset is split into training and testing sets to evaluate the performance of the model.

Model Selection: In this step, different machine learning algorithms such as logistic regression, decision tree, random forest, and XGBoost are trained on the training set and evaluated on the testing set. The model with the highest accuracy and F1-score is selected for further analysis.

Model Tuning: In this step, hyperparameters of the selected machine learning model are tuned using techniques such as grid search and randomized search to improve the model's performance.

Model Deployment: In this step, the final model is deployed and used to classify new URLs as phishing or legitimate. The model's performance is evaluated on a validation set and compared to the testing set to ensure that the model's performance remains consistent.

Literature Review

Phishing is a type of cyber attack in which an attacker tricks a user into revealing sensitive information, such as login credentials, by disguising themselves as a trustworthy source. Websites are one of the most common forms of phishing attacks. To prevent such attacks, machine learning algorithms can be used to detect and thus prevent them. In this literature review, we will explore various repositories, frameworks, and research articles related to phishing website detection using machine learning algorithms.

Detecting Phising attacks has been successful with machine learning and there has been some fundamental researches regarding this.

1. A Comprehensive Study of Phishing Attacks
Link: https://www.semanticscholar.org/paper/A-Comprehensive-Study-of-Phishing-Attacks-Banu-Banu/2bf12ff75150903efee426f23035c94d599597ae

In the paper "A Comprehensive Study of Phishing Attacks," the authors used various techniques to detect phishing attacks. They evaluated the effectiveness of different anti-phishing solutions, including blacklists, heuristics, and machine learning algorithms. They found that machine learning algorithms, such as decision trees, random forests, and support vector machines, were the most effective approach for detecting phishing attacks, with accuracy rates of up to 99

2. Phishing Detection Based on Decision Tree Algorithm
Link: https://www.learntechlib.org/p/216410/
The paper "Phishing Detection Based on Decision Tree Algorithm" proposes a machine learning-based approach to detecting phishing websites that bypass traditional signature-based detection systems. The authors use a decision tree algorithm to classify websites as phishing or legitimate based on features extracted from the website's HTML source code. The paper reports an accuracy of 94.5

3. Phishing website detection based on effective machine learning approach
Link: https://www.tandfonline.com/doi/abs/10.1080/23742917.2020.1813396
The paper proposes a novel machine learning approach for detecting phishing websites. The authors address the problem of the increasing number of sophisticated phishing attacks that bypass traditional detection methods. The proposed

approach uses a hybrid feature selection technique along with machine learning algorithms to classify websites as phishing or legitimate with an accuracy of up to 98.5

4. Feature Selection Approach for Phishing Detection Based on Machine Learning
Link: $https://link.springer.com/chapter/10.1007/978-3-030-95918-0_7$

The paper proposes a feature selection approach for detecting phishing websites using machine learning. The authors address the problem of the increasing sophistication of phishing attacks and the need for effective and efficient detection methods.The proposed approach uses a feature selection technique to reduce the number of features used by the machine learning algorithm, thus improving the performance and reducing the computation time. The authors compare the performance of their approach with other feature selection techniques and report an accuracy of up to 98.46

5. Phishing Detection Using Machine Learning Techniques
Link: https://paperswithcode.com/paper/phishing-detection-using-machine-learning
The paper proposes a machine learning-based approach to detect phishing websites that use sophisticated techniques to evade detection. The approach uses a combination of feature selection techniques and machine learning algorithms to classify websites as phishing or legitimate, achieving an accuracy of up to 98.8

6. Machine learning based phishing detection from URLs
Link: https://www.sciencedirect.com/science/article/abs/pii/S0957417418306067
The paper above proposes a machine learning approach to detect phishing URLs by extracting features and selecting important ones to improve the accuracy of classification. The study addresses the need for effective and efficient phishing detection methods due to the increasing sophistication of phishing attacks. The authors report an accuracy of up to 98.5

7. A machine learning based approach for phishing detection using hyperlinks information
Link: https://link.springer.com/article/10.1007/s12652-018-0798-z

The paper proposes a machine learning approach to detect phishing using hyperlink information. It addresses the increasing sophistication of phishing attacks and the need for efficient detection methods. The proposed approach

uses machine learning algorithms to classify emails as phishing or legitimate based on hyperlink information, with an accuracy of up to 99.1

8. PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning
Link: https://link.springer.com/chapter/10.1007/978-981-10-8536-9$_4$4

The paper "PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning" proposes a machine learning approach for detecting phishing URLs, addressing the problem of the increasing sophistication of phishing attacks. The approach uses a machine learning algorithm to classify URLs based on extracted features and achieves an accuracy of up to 97.68

Respiratories:

1. Phishing Detection Using Machine Learning Techniques
Link: $https://github.com/fafal-abnir/phishing_detection$

The project aims to develop a machine learning model to detect phishing websites using a dataset of 10,000 URLs and the Random Forest algorithm. The model extracts features such as domain age, length, special characters, and SSL certificate from the URLs. The project's scope is to provide a solution to the problem of online security by detecting phishing URLs using machine learning, which could be valuable for both industry and academic research.

2. Phishing Website Detection by Machine Learning Techniques
Link: https://github.com/shreyagopal/Phishing-Website-Detection-by-Machine-Learning-Techniques

The "Phishing Website Detection by Machine Learning Techniques" repository detects phishing websites using a dataset of 11,055 URLs. The dataset contains both legitimate and phishing URLs collected from different sources. Several machine learning models such as logistic regression, decision trees, k-nearest neighbors, and support vector machines are used with feature selection techniques to improve model performance. The scikit-learn library is used for model building and evaluation. The repository reports high accuracy of 96.18

3. Phishing Attack Domain Detection
Link: https://github.com/deepeshdm/Phishing-Attack-Domain-Detection

The "Phishing Attack Domain Detection" repository aims to detect phishing domains using machine learning techniques. The dataset used in the project contains 10,000 domains, half of which are legitimate and the other half are phishing. The repository uses machine learning models such as random forest, decision trees, k-nearest neighbors, and support vector machines. Feature extraction and selection techniques such as the Levenshtein distance metric, domain entropy, and n-grams are used. The random forest model with domain entropy and n-grams feature extraction achieved the highest accuracy of 98.6

The dataset that I will be using is textual and is in csv file. I aim to implement various Ml algorthm on the dataset and compare the efficiency method and based on the model preprocessing of the data will also be done.

Result

The results indicated that the Random Forest algorithm performed the best in identifying phishing websites, achieving an accuracy score of 98.85

Furthermore, a feature selection analysis ahs been conducted to determine which website parameters contributed the most to identifying phishing websites. The results indicated that the presence of the "@" symbol in the website's URL was the most significant indicator of a phishing website. Other features such as the length of the URL, the use of redirection, and the presence of a hyphen in the URL were also found to be relevant.

Overall, the results of the project demonstrated the potential of machine learning algorithms in detecting and preventing phishing attacks. By utilizing these algorithms and identifying crucial website parameters, it's possible to develop effective anti-phishing measures that can protect users from falling victim to these attacks.