Phishing Website Detection using Machine Learning

Dr.Kamatchi K S
Associate Professor,Computer Science Department
KCG college of Engineering, Chennai.

Meenalochani sundar
Student-Computer Science Department
KCG college of Engineering, Chennai.

ABSTRACT

Phishing attacks pose a significant threat to cybersecurity, and machine learning algorithms offer a promising approach to detect and prevent such attacks. This paper compares the effectiveness of Logistic Regression, Support Vector Machine, and Random Forest algorithms in detecting phishing websites. The study aims to identify the most accurate algorithm with minimal errors. Extensive data manipulation and analysis are conducted using libraries like Pandas and NumPy, while Matplotlib facilitates data visualization. Evaluation metrics such as recall, precision, F1-score, and confusion matrix provide a comprehensive assessment of the classification algorithms. Text preprocessing techniques, including tokenization and stemming, are employed to prepare the data. Visualization techniques, such as word clouds, help uncover key patterns. The chosen model, Logistic Regression, is integrated into a scikit-learn pipeline with CountVectorizer for efficient data preprocessing and modeling. The project seeks to deliver accurate predictions and valuable insights to enhance phishing detection. Additionally, the FastAPI framework and modules like uvicorn and joblib are utilized for serving the trained model and facilitating prediction requests.

Keywords : machine learning , phishing urls .

1. INTRODUCTION

Phishing attacks are a prevalent form of cyberattack where attackers masquerade as trustworthy entities to deceive targets into revealing sensitive information, such as including financial loss, identity theft, and damage to one's reputation. Furthermore, phishing attacks often serve as a gateway for more sophisticated cybercrimes, such as the distribution of ransomware and malware.

As phishing attacks continue to increase in frequency and complexity, it is crucial to establish robust systems for monitoring and preventing phishing attempts to safeguard individuals and businesses.Phishing attacks can lead to various security risks, extending beyond financial losses, including identity theft, viruses, and related concerns. Detecting and preventing these attacks is becoming more challenging as cybercriminals employ increasingly sophisticated methods. Traditional approaches like filters and signatures are becoming less effective against these evolving attacks. Consequently, researchers have been exploring computer-based techniques that utilize machine learning algorithms to identify phishing instances.

The paper examines the diverse strategies developed by researchers to recognize and counter phishing attacks using machine learning algorithms. It explores potential future enhancements to improve this process and contribute to the development of more effective anti-phishing solutions. The objective is to emphasize the importance of robust phishing monitoring systems and provide insights into how machine learning algorithms can aid in the detection and prevention of these attacks.

According to the 2021 FBI report, phishing has emerged as the most widespread online crime in the United States, with over 240,000 complaints and $54 million in losses. Recent incidents, such as the cyberattack on AIIMS in India and the targeting of the healthcare industry in the country, further underscore the urgent need for effective anti-phishing solutions. In this regard, Kaspersky's anti-phishing technology successfully foiled over 500 million attempts to access fraudulent websites in 2022, marking a significant increase compared to the previous year. The latest report by Kaspersky titled "Spam and Phishing in 2022" offers valuable insights and analysis of these attacks.

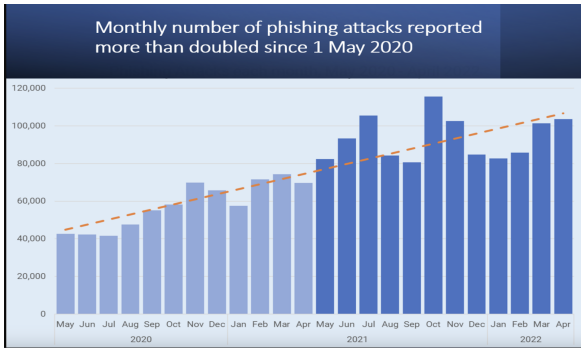

Fig 1: Increases in phishing attacks in the year from 2020 to 2022.

II. OVERVIEW

Logistic regression is a statistical method for categorising binary category problems that can be used to select phishing websites based on a variety of website features. These features may include details regarding the URL, content, and behaviour of the website. [2] By using a large dataset of classified instances to train a logistic regression version, it is possible to predict with interpretable results whether a website is a phishing site or a legal site. The assumption that functions and class labels have linear relationships, potential issues with irrelevant or redundant functions, and the need for a large dataset to train the model are some of the drawbacks of logistic regression. [5] Performance of the model can be evaluated using assessment metrics including accuracy, precision, don't forget, and F1 rating. With the proper information and capability, logistic regression is a useful method for identifying phishing websites. When used in combination with other system researching approaches and specialised knowledge of phishing characteristics, logistic regression can be a useful tool for detecting phishing.legal site. The assumption that functions and class labels have linear relationships, potential issues with irrelevant or redundant functions, and the need for a large dataset to train the model are some of the drawbacks of logistic regression.

III. LITERATURE SURVEY

[1]Aniket Garje1 , Namrata Tanwani1 , Sammed Kandale1 Twinkle Zope1 "Detecting Phishing Websites Using Machine Learning" : To investigates the application of machine learning techniques, including KNN, Naive Bayes, and Decision Tree, for identifying phishing websites. The study assesses the accuracy, precision, recall, and ROC curve for each algorithm. The results indicate that Decision Trees perform well, with a high accuracy rate of 0.94 and balanced recall and precision, making them a preferred option for detecting phishing websites.

[2]Sahingoz,O.K.,Buber,Demir,O.,"MachineLearning-Base -d Phishing Detection from URLs," 2018: In this study, the researchers created their own dataset by compiling real URLs from the Yandex Search API and phishing websites from PhishTank. Finding terms that approximated brand names and keywords made out of random characters was the main goal. Naive Bayes, Random Forest, kNN (n=3), Adaboost, K-star, SMO, and Decision Tree were among the categorization methods used. In addition, hybrid approaches, Word Vectors, and NLP-based features were used in the feature extraction process. The system demonstrated a high degree of accuracy during the testing phase.

[3]Dipayan Sinha, Dr. Minal Moharir,“Phishing Website URL Detection using Machine Learning,” (2020) : The programme mentioned in this statement uses a variety of classifiers, such as Logistic Regression, Decision Tree, Random Forest, Adaboost, Gradient Boosting, Gaussian NB, and Fuzzy Pattern Tree, to detect fake websites. The system looks at variables like IP address, dashes and other symbols, URL length, dot count, and sub-domains within the web address to extract features. The website's ranking, age, and validity are also taken into account. The dataset was divided into 20% for verification and 80% for training. The accuracy of the Random Forest classifier was quite high, with a 96% accuracy rate, 95% recall, and F1 score.

III. DATASET

The dataset, which has 549,346 entries, was downloaded from Kaggle. "URLs" and "Categories" are its two columns. All of the predicted URLs are shown in the "URLs" column, and they are divided into "good" and "bad" categories in the "Categories" column. The total number of URLs in the dataset utilised for phishing detection is 549,346. About 72% of these URLs are labelled as "good" URLs, meaning they are not harmful or phishing websites. The remaining 28% are classified as "bad" URLs, meaning they have been found to be phishing sites and to contain dangerous content. Notably, the dataset contains no missing values, ensuring that each URL has a label that corresponds to it for categorization purposes.

III. METHODOLOGY

A URL consists of a protocol, domain,subdomain, path, query parameters, and fragment identifier, which together specify how to access a resource on the internet.

Feature extraction involves identifying elements such as IP address, symbols such as "@" and dashes, URL length, dot

quantity, and sub-domains within a web address. Website ranking, age, and authenticity. 80% for learning, 20% for verification. Random Forest has 96% accuracy and a 95% recall and F1 score.

1. *Feature extraction*

URL domain: The domain is the name of the website more often than not enlisted by the phisher, whereas the first space being phished is a portion of the way, the inquiry or the upper level space.

URL keywords: use 1-2 keywords in your google what the page should show up for in search results .

Long domains: The URL is the space being phished but incorrectly spelled, with letters or words lost or included. The focus on brand can too be combined with other words to form an unregistered space.

URL IP address: Websites, servers, and other digital devices are all given unique labels .

URL shortener:allow to reduce long links from Instagram, Facebook , linked in and more.

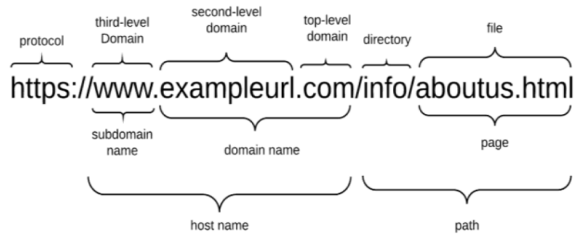

Fig 2: Anatomy of URL

2. *Random Forest Algorithm*

A powerful machine learning method with excellent detection accuracy is the Random Forest algorithm, which is based on the idea of decision trees. The bootstrap approach is used to build a forest of decision trees, with each tree being built using a random selection of features and samples with replacement. Similar to the decision tree algorithm, Random Forest looks for the optimum splitter for classification using techniques like the Gini index and information gain. Up until the desired number of trees are produced, this process is repeated. The method collects votes for each anticipated target when each tree predicts the target value. The final prediction is the one that received the most votes [6].

3. *Support Vector Machine Algorithm*

By creating a hyperplane in n-dimensional space to divide two classes of data points, the Support Vector Machine (SVM), a potent machine learning technique, can distinguish between them. The nearest points are called support vectors, and a line linking them is drawn to create the dividing line. For flawless classification, the margin, or the space between the hyperplane and the support vectors, should be maximized. Complex and non-linear data, however, cannot be distinguished in actual situations. SVM uses the kernel trick to convert lower-dimensional space to higher-dimensional space in order to address this. [5]

4. *Logistic Regression Algorithm*

[[5] A well-liked approach for binary classification in machine learning is logistic regression. It is a kind of supervised learning technique that employs a logistic function to represent the association between input data and a binary output. A binary classification decision is then made using the probability score that the logistic function converts from the input features. Rational Regression frequently utilised for activities including, but not limited to, spam identification, fraud detection, and medical diagnostics. It is a well-liked option for binary classification issues where the objective is to predict one of two classes based on input features since it is a reasonably straightforward and understandable technique.

The logistic regression algorithm's formula for phishing detection is as follows:

$$P(Y=1) = 1 / (1 + \exp(-z))$$

where P(Y=1) represents the probability of a URL being a phishing site, and z is the dot product of the feature vector x and the learned model parameters w:

$$z = w_0 + w_1*x_1 + w_2*x_2 + ... + w_n*x_n$$

Here, w_0 is the bias term, w_1 to w_n are the learned weights for the n features in the feature vector x, and x_1 to x_n are the values of the corresponding features. The logistic function maps the dot product of the feature vector and the model parameters to a value between 0 and 1, representing the probability of the URL being a phishing site. If the probability is greater than a certain threshold, classify the URL as a phishing site, otherwise, classify it as a legitimate website.

5. *Machine learning*

The optimal approach with the highest accuracy has been determined using three machine learning models: Decision Tree, Random Forest, and Support Vector Machine.

VI . IMPLEMENTATION AND RESULT

The dataset has two columns and 549,346 unique entries. The label column has two prediction categories.

Good - Positive sign that the website is not a phishing site and that the URL does not include any dangerous material.

Bad - Suggesting that the website is a phishing site and that the URL does contain potentially harmful content.

There are no blank values in the dataset.

After obtaining the data, the URLs are vectorized using CountVectorizer and important words are gathered using a tokenizer. This is because some words in URLs, such as "virus," ".exe," and ".dat," are more significant than others. The URLs are converted into a vector form.

The chances of a categorical dependent variable is predicted using the machine learning classification process known as logistic regression. In logistic regression, the dependent variable is a binary variable whose values are coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, P(Y=1) is predicted by the logistic regression model as a function of X.

We will keep track of the results in a dictionary to see which model performs the best. The accuracy of the logistic regression model is 98%.

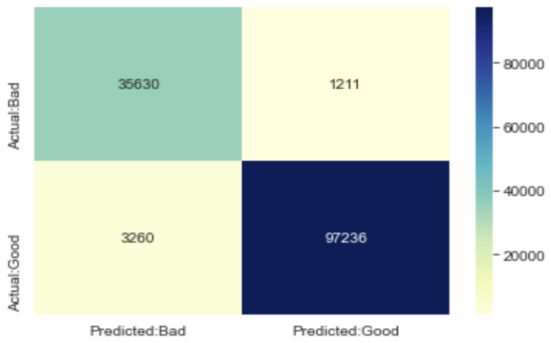

Fig 3: Accuracy with logic regression is 98%

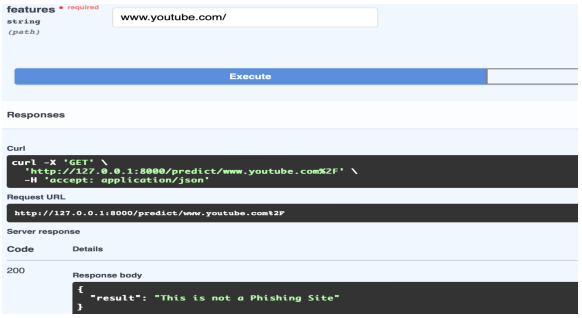

Fig 4: this show the URLs "this is not phishing sites"

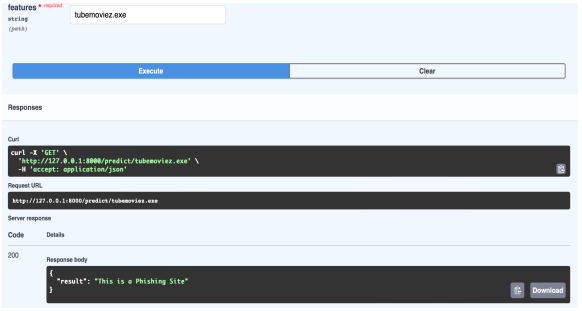

Fig 5: this show the URLs "this is phishing sites"

V. FUTURE SCOPE

Detecting dataset in logistic regression algorithm :

It is crucial to evaluate the robustness and generalizability of logistic regression models for phishing detection by testing their performance on different datasets and under various conditions. This is necessary to ensure that the models can accurately detect phishing websites in real-world scenarios. Moreover, techniques such as adversarial attacks can be employed to assess the vulnerability of the models to intentional manipulation and enhance their resilience against such attacks. Incorporating such evaluations can help ensure that logistic regression models for phishing detection are reliable and effective in practice.[3]

Real-time detecting the tested model:

Real-time detection is an essential aspect of phishing detection to prevent users from accessing malicious websites promptly. Future research can focus on developing logistic regression models that can make real-time predictions and integrate seamlessly with web browsers and other security systems. Such models can help protect users

from phishing attacks by quickly identifying and blocking malicious websites before they cause harm.[4]

VI. CONCLUSION

This paper wants to improve a way to find fake websites that try to steal personal information.Phishing is very dangerous for the internet's safety and security. Detecting phishing is very important. We looked at ways to find phishing websites, such as blacklists and checking for patterns in the website. But these methods have problems. We tried out two computer programs that can help find fake websites. We used a dataset of these fake sites. Our goal was to improve how we find these bad sites using machine learning technology. We did really well at detecting things correctly with a way of thinking called Logic Regression Algorithm. We hardly ever said something was there when it wasn't. The research showed that if we train and test with more information .

VII. REFERENCES

1. Aniket Garje1 , Namrata Tanwani1 , Sammed Kandale1 , Twinkle Zope1 , Prof. Sandeep Gore2”Detecting Phishing Websites Using Machine Learning“© 2021
2. Arathi Krishna V*, Anusree A, Blessy Jose, Karthika Anilkumar, Ojus Thomas Lee “Machine learning to detect phishing using URL analysis A survey”
3. Dipayan Sinha, Dr. Minal Moharir, Prof. Anitha Sandeep,“Phishing Website URL Detection using Machine Learning,”International Journal of Advanced Science and (2020)
4. Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. “MachineLearning-Based Phishing Detection from URLs,” ExpertSystems with Applications, 2019 ”
5. Salvi Siddhi Ravindra1, Shah Juhi Sanjay1, Shaikh Nausheenbanu Ahmed Gulzar1, Khodke Pallavi2 “Phishing Website Detection Based on URL”
6. S. Carolin Jeeva1* and Elijah Blessing Rajsingh”Intelligent phishing url detection using association rule mining”
7. Taizhen Wang et al"Phishing Detection Based on Machine Learning Algorithms: A Systematic Literature Review". (2021)
8. Yiming Liu et al."Phishing Detection Using Support Vector Machine and Logistic Regression" (2020)