IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

Phishing URL's Prediction

Akash Soni
Department of Information Technology
Madhav Institute of Technology and Science, Gwalior, India

Abstract: *Identity theft is the act of stealing our precious, personal, sensitive data such as the information we use to access services, Available throughout the cyberspace. Recognizing this rapidly growing cybercrime and its negative impact on businesses including individual users, it is now necessary for organizations and individuals around the world to successfully predict a sensitive identity theft website and separate it from legal ones. The purpose of this project is to successfully predict criminal websites for stealing sensitive information so that users can benefit from this project and prevent them from being caught. In this project, machine learning methods are used to predict. Data mining is used worldwide almost every face of the community i.e. business associations, govt. organizations, and other types of data collectors to extract information from collected data.*

Keywords: Machine Learning, Data Mining, URL's, Phishing

I. INTRODUCTION

Identity theft is a form of fraud in which an attacker attempts to read sensitive information such as login details or account information by sending it as a reputable business or person via email or other communication channels.

Often the victim receives a message that appears to have been sent by a known contact or organization. The message contains malicious software targeting the user's computer or contains links to target victims to malicious websites to trick them into disclosing personal and financial information, such as passwords, account IDs or credit card details.

Theft of sensitive information is popular with attackers, as it is easier to trick someone into clicking on a malicious link that appears to be legitimate than trying to hack computer security systems. Malicious links within the message body are designed to make it appear that they are going to a corrupt organization using that organization's logos and other official content.

To obtain confidential data, criminals create unauthorized analogy of the actual website and email, usually from a financial institution or another organization that holds financial data. This email address is provided using official company logos and slogans. HTML design and layout allows for the copying of images or the entire website. Also, it is one of the fastest growing features of the Internet as a means of communication, and it allows for the misuse of products, trademarks and other company identifiers that customers rely on as security mechanisms. To alert users, Phisher sends "compiled" emails to as many people as possible. When these emails are opened, customers are often diverted from legitimate business to fraudulent websites.The objectives of the study are as follows:

- Creating a new way to find malicious URLs and notifying users.
- Using ML techniques in the proposed method to analyze real-time URLs and produce effective results.

To implement this project, which is a standard ML method with the ability to handle large amounts of data

1.1: DATASET

Content: The data contains 5,49,346 entries. There are two columns.

The label column is a 2-step prediction column

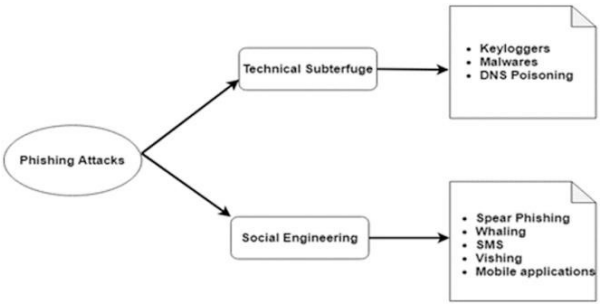
- **Good** - which means that URLs do not contain malicious content and that this site is not a criminal site for stealing sensitive information.
- **Bad** - meaning that the URLs contain malicious content and that this site is a site for theft of sensitive information.

There is no missing value in the database.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

1.2: PHISHING TECHNIQUES

Here we discuss some well known phishing techniques which is used to deceive people. Crime-stealing websites are sensitive to the organization and the individual because of their similarity to official websites. Figure 1 shows the many types of criminal attacks of sensitive information theft. Technological fraud refers to attacks that include Keylogging, DNS poisoning, and Malware. In this attack, the attacker aims to gain access to the tool / strategy. On the one hand, users believe in the network and on the other hand, the network is vulnerable to attackers. Social engineering attacks include identity theft, Whaling, SMS, Vishing, and mobile applications. In this attack, the attackers target a group of people or an organization and trick them into using the criminal URL to steal sensitive information. Despite these attacks, many new attacks are emerging as technology evolves constantly.



1.3: PHISHING DETECTION APPROACHES

Various measures have been taken to curb the theft of sensitive information attack with each level of attack flow. Some of them methods need to train users to be ready for the future attacks and some of them are automatic and alert user. These methods are as follows.

- User training
- Software detection

User training:

Educating the company's users and employees and warning them with regard to the crime of theft of sensitive information contributes to the prevention of crime of identity theft to attack. Many training methods have been suggested users. Many studies have concluded that it has far-reaching effects how to help users differentiate between the crime of stealing sensitive information as well official websites for interactive teaching. Although user training is an effective but human error they still exist and people tend to forget their training. Training and it requires significant time and not much notified by non-technical users

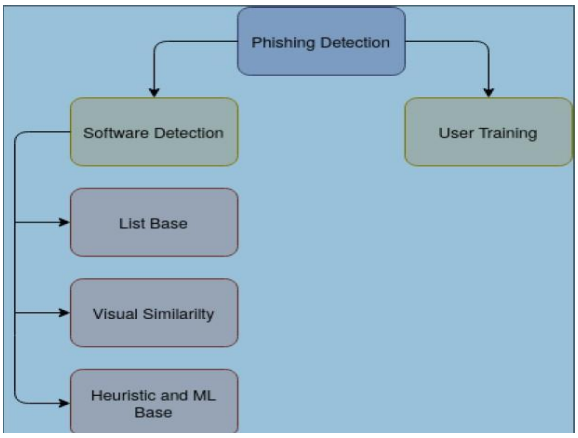
Software detection:

Although user training can prevent criminal attacks from stealing sensitive information however we are attacked daily by hundreds of websites so applying our training to each website is difficult and sometimes unrealistic work. One way finding criminal websites to steal sensitive information using software. I software can analyze many features such as the content of website, email message, URL, and many other features before makes its final decision more reliable than the people. Several software options are proposed to detect the theft of sensitive information

Types of approaches of software detection:

1. List-Base approach
2. Visual similarity-Base approach
3. Machine learning based

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252



II. MACHINE LEARNING APPROACHES

Machine learning offers simple and effective methods data analysis. Show realistic results in real time recent separation problems. The main benefit of machine learning is the ability to create flexible model certain activities such as the detection of identity theft. As the theft of sensitive information is separation problem, machine learning models can be used as a powerful tool. Machine learning models can become familiar changes quickly to identify patterns of fraudulent activity which helps to create a learning-based identification system. Most of the machine learning models mentioned here is considered a surveillance machine, This is where The algorithm attempts to read a function that displays input in output based on the input-output pair model. Considered a job from a training data labeled that includes a training set examples. We present the machine learning methods we have used in our study

2.1 Logistic Regression

Logistic Regression is a classification algorithm used give visuals to a set of different classes. Unlike the line retractable output of continuous numerical values, Logistic Undo reverses the output using a logistic sigmoid the task of restoring the number of opportunities that may be possible the map is divided into two or more separate classes. Backlash works best when the data relationship is almost equal despite the fact that there are complex indirect relationships in between variable, it has poor performance. Other than that, it needs more mathematical guessing before applying other techniques.

2.2 Decision Tree

The categories of decision trees are used as the most popular categories technology. Decision tree is a tree- like tree structure where the internal node represents a feature or attribute, i the branch represents the law of decision, and each area of the leaf represents the result. The highest node in the decision tree is known as root node. Learns to distinguish based on attribute value. It separates the tree in a recurring form called recursive to separate. This particular feature gives the tree category a high resolution to deal with a variety of data sets, however numerical or category data. Also, pruning trees are ideal to deal with the indirect relationship between factors and classes. Generally, the contamination function is determined by testing the separation quality of each node, as well as the Gini Variety The index is used as a known indicator for full functionality. In fact, the tree decides and conditions in the sense that can easily model indirect or unfamiliar relationships. It can translate interactions between forecasters. It could be too translated very well because of its binary structure. However, decision medicine has a variety of problems that are often overused data. Alternatively, update the decision tree with new samples hard.

2.3 MultinomialNB

The Multinomial Naive Bayes algorithm is a possible learning method widely used in Natural Language Processing (NLP). The algorithm is based on a Bayes perspective and predicts text markings such as a piece of email or a newspaper article. It calculates the probability that each sample mark is given and gives the mark with the highest probability as output. The Naive Bayes classifier is a collection of multiple algorithms in which all algorithms share one common goal,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

and that each distinguished feature is not related to any other element. The presence or absence of an element does not affect the presence or absence of another element.

2.4 Random Forest

A random forest, as its name implies, contains a large amount of decision-making trees that work as a team to decide output. Each tree in a random forest defines a class prediction, and the result will be the most predicted category in the middle of a tree decision. The reason for this is amazing the effect from Random Forest is because the trees protect each other from individual mistakes. Although some trees may predicting the wrong answer, many other trees will fix it final prediction, so that as a group the trees can move right guidance. Random Forests benefit from the reduction of overgrazing by combining many weak students who do not succeed well because they use only a subset of all training Trees for the Random Forest can handle data set. Also, during the construction of the forest, they act impartially standard error rate. Otherwise, they can equate well-lost data. The main drawback of Random Forests lack of fertility because of the forest process construction is unplanned. Otherwise, it is difficult to translate the final model and the following results, because it involves many trees independent decisions

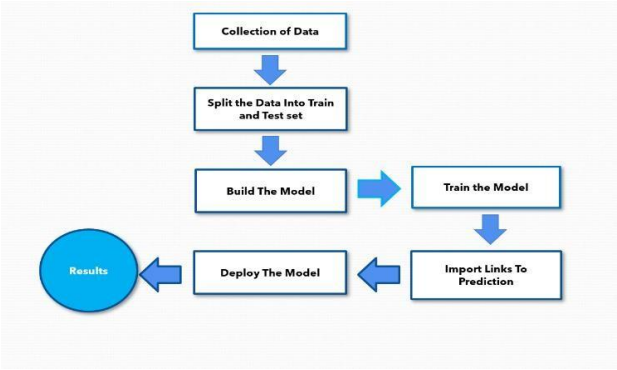
2.5 Ada-Boost

From other factors, Ada-boost is similar to Random Forest, i The division of Ada-Boost as Random Forest groups is weak split models to form a solid phase. One the model may misinterpret things. But if we put it together a few dividers by selecting a set of samples for each repetition and given enough weight in the final vote, it would be good according to the general categories. Trees are created in sequence as weak students and correcting incorrectly speculated samples by giving them more weight after each cycle forecast. The model learns from past mistakes. I Final forecast is a weighted (or rated) majority vote median in case of retreat problems). Briefly Ada-Boost the algorithm is repeated by selecting a training set based on the accuracy of previous training. The weight of each phase training in each repetition depends on the accuracy obtained from past.

III. IMPLEMENTATION

- Collect dataset containing phishing and legitimate websites from the open source platforms.
- Write a code to extract the required features from the URL database.
- Analyze and preprocess the dataset by using EDA techniques.
- Divide the dataset into training and testing sets.
- Write a code for displaying the evaluation result considering accuracy metrics.
- Compare the obtained results for trained models and specify which is better.

3.1 Proposed Approach



3.2 Feature Extract

Collected data from database undergo feature extraction form that urls into words.Used CountVectorizer and gather words using tokenizer, since there are words in urls that are more important than other words e.g 'virus', '.exe' ,'.dat' etc. Lets

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

convert the URLs into a vector form.
Then these words stemmed. Then good and bad urls sent

3.3 Methods Used

From the above Machine learning methods. I implement two of them in this project to find out which one is better in case of accuracy and test time

Evaluation Metrics: For testing the results obtained, we used 3 parameters: Accuracy, Recall and False Positive Rate (FPR).

Accuracy: Estimated number of correct predictions and total number of input samples. Since purpose requires multiple URLs to be categorized correctly, this is why high accuracy is one of the metrics.

Recall: Average number of true points and total number of predicted points. Since we want websites that are predicted to be accurate, only legitimate, high memory is required.

False Positive Rate (FPR): Estimated number of incorrectly identified samples as positive to the total number of incorrect samples. The requirement is to reduce the number of criminal websites to steal sensitive information that has been identified as legal as it could result in significant losses to the website visitor. So low FPR is one of the metrics used.

3.4 Fit Model

Logistic Regression

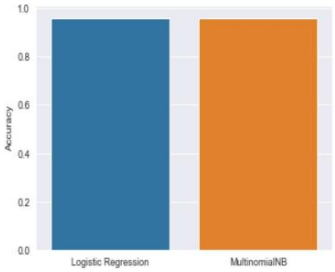
Logistic Regression is a classification algorithm used give visuals to a set of different classes. Unlike the line retractable output of continuous numerical values, Logistic Undo reverses the output using a logistic sigmoid the task of restoring the number of opportunities that may be possible the map is divided into two or more separate classes. Backlash works best when the data relationship is almost equal despite the fact that there are complex indirect relationships in between variable, it has poor performance. Other than that, it needs more mathematical guessing before applying other techniques.

MultinomialNB

The Multinomial Naive Bayes algorithm is a possible learning method widely used in Natural Language Processing (NLP). The algorithm is based on a Bayes perspective and predicts text markings such as a piece of email or a newspaper article. It calculates the probability that each sample mark is given and gives the mark with the highest probability as output. The Naive Bayes classifier is a collection of multiple algorithms in which all algorithms share one common goal, and that each distinguished feature is not related to any other element. The presence or absence of an element does not affect the presence or absence of another element

IV. RESULT

On the basis of above data. Logistic regression performing well with high accuracy, high recall and low FPR. So for this project I am going with logistic regression method because is test accuracy is higher which is 96%.



• So, Logistic Regression is the best fit model, Now we make sklearn pipeline using Logistic Regression

4.1 Deployment of ML Model Using FastAPI

A. FastAPI

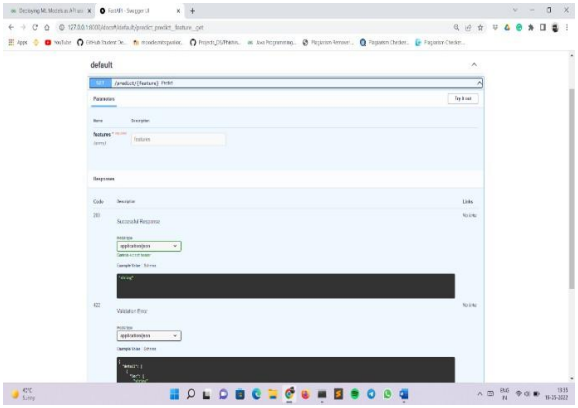
FastAPI is a Python framework and set of tools that enables developers to use a REST interface to call commonly used functions to implement applications. It is accessed through a REST API to call common building blocks for an app. In

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

this example, the author uses FastAPI to create accounts, login, and authenticate.

B. Model Deployment

To test our API we will be using the Swagger UI now to achieve what you will just need to add / document at the end of your path. So go to http: //127.0.0.1: 8000 / docs. You should also see the following output:



V. CONCLUSION

The proposed study has emphasized the criminal approach to identity theft in the context of segregation, where the criminal identity theft website is considered to include the automatic classification of websites into a predetermined set of class values based on fewer features and class variability. ML-based criminal identity theft strategies relied on website performance to collect information that could assist in classifying websites in order to detect criminal sites to steal sensitive information. The problem of identity theft cannot be eliminated, however it can be reduced by fighting it in two ways, developing anti-crime information systems and strategies and informing the public how fraudulent criminal websites can be identified and identified. To combat the recurring and complexity of cybercrime, ML strategies to combat identity theft are essential. This project aims to improve access to crime hacking websites for sensitive information using machine learning technology. We found 96 % detection accuracy using a logistic regression algorithm with very low false value. And the result shows that the dividers provide the best performance when we use additional data as training data. In the future, integrated technologies will be used to detect criminal websites to steal sensitive information more accurately, in which case the logistic regression algorithm of machine learning technology and method of restricted listing will be used.

VI. FUTURE SCOPE

The results encourage future activities to add additional features to data, which can improve the performance of these models, it can cope ML model with strategies for detecting identity theft, for example List-Base methods to get better performance. In addition, we will explore suggest and develop a new way of releasing new features from the website to keep up with new crime scams to steal sensitive information to attack. In the future, integrated technologies will be used to detect criminal websites to steal sensitive information more accurately, in which case the random forest algorithm of machine learning technology and method of restricted listing will be used.

ACKNOWLEDGEMENT

The full semester project has proved to be pivotal to my career. I am thankful to my institute, **Madhav Institute of Technology and Science** to allow me to continue my disciplinary/interdisciplinary project as a curriculum requirement, under the provisions of the Flexible Curriculum Scheme (based on the AICTE Model Curriculum 2018), approved by the Academic Council of the institute. I extend my gratitude to the Director of the institute, **Dr. R. K. Pandit** and Dean Academics, **Dr. Manjaree Pandit** for this.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

I would sincerely like to thank my department, **Department of Information Technology,** for allowing me to explore this project. I humbly thank **Dr. Akhilesh Tiwari**, Professor and Head, Department of Information Technology, for his continued support during the course of this engagement, which eased the process and formalities involved. I am sincerely thankful to my faculty mentors. I am grateful to the guidance of **Abhilash Sonkar**, Assistant Professor, Information Technology, for his continued support and guidance throughout the project. I am also very thankful to the faculty and staff of the department.

REFERENCES

[1]. Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey IEEE, and Andrew Jones, 2013
[2]. http://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing
[3]. A. Alswailem, B. Alabdullah, N. Alrumayh, and A. Alsedrani. Detecting phishing websites using machine learning. In 2019 2nd International Conference on Computer Applications Information Security (ICCAIS), pages 1–6, 2019.
[4]. M E Pratiwi, T A Lorosae, and F W Wibowo. Phishing site detection analysis using artificial neural network. Journal of Physics: Conference Series, 1140:012048, dec 2018.
[5]. www.phishtank.com
[6]. https://archive.ics.uci.edu/ml/datasets/Phishing+Websites
[7]. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/
[8]. https://stackoverflow.com/questions/tagged/phishing
[9]. https://phishstats.info/