# Phishing URL Detection using Machine Learning

**Deepak Chandradev Prajapati and Nikhil Chandrajeet Yadav**
Master of Computer Applications
Institute of Distance and Open Learning, Mumbai, Maharashtra, India
deepakprajapati78918@gmail.com and Nikhil.ncsy@gmail.com

**Abstract***: Phishing attack is an easy way to obtain sensitive information from innocent users and also there are many tools and techniques available to do so. The main aim of the attackers is to get the critical information like username, password, bank account details and etc. and by this they can harm the innocent users. As a security perspective there should be a system or tools that would be able to overcome such issues. So this paper deals with machine learning technologies for the detection of phishing URLs by analyzing and extracting different types of features of legitimate and phishing URLs. Various machine learning algorithms like Decision Tree, Random Forest and Support Vector Machine are used to detect phishing websites. Aim of this paper is to detect phishing urls as well as narrow down to best machine learning algorithm by comparing the accuracy rate, false positive and false negative rate of each algorithm.*

**Keywords:** Machine learning, Phishing URL Detection Legitimate websites and Phishing websites

## I. INTRODUCTION

In recent years, there have been many advancements in Internet and cloud technologies that have a significant increase in electronic trading where customers make transactions by purchasing or buying items. This growth leads to unauthorized access to user's sensitive information and harm the resources or systems of an enterprise or an individual. Phishing is one of the simplest and familiar attack that exploits the user's weakness and steal critical information.

Nowadays phishing becomes an important sector of concern for security researchers because it is not difficult to create fakes websites that will looks as real as legitimate websites. A security experts or well knowledgeable person can identify such fake websites but it's not that easy for all users to identify the fake websites and such users become the victim of phishing attack. Main aim of the Phishing attackers is to steal banks account credentials.

Phishing attacks are becoming successful due to lack of user awareness. Since phishing attack exploits the weakness found in users, it is very difficult to mitigate them but it is very important to enhance the tools or the phishing detection techniques. Machine learning technologies consists of many algorithms which requires past data to make a decision or prediction on future data using these techniques, algorithms will analyze various blacklisted and legitimate URLs and their feature to accurately detect the phishing websites including zero-hours Phishing websites.

## II. METHODOLOGY

### DATASETS

URLs of legitimate websites can be collected from www.alexa.com and the URLs of Phishing websites can be collected from www.phishtank.com. After the collection of websites, we will label the legitimate websites as "0" and "1" for phishing websites. The main working of this system is based on Feature Extraction. Where we will implement a python program to extract features from URLs.

### FEATURE EXTRACTION:

**Presence of IP address in URL:**

Most of the legitimate websites do not use IP address as an URL to download a webpage. If the IP address is present in the URL, then it will be an indications that attacker is trying to steal sensitive information. So, if the IP address is present in URL the feature set to "1" else set to "0".

DOI: 10.48175/IJARSCT-15248

**Presence of @symbol in the URL:**
Attackers uses such types of special symbol @ in the URL that leads the browsers to ignore everything preceding the "@" symbol and the real address often follows the "@" symbol. So if the @ is present in URL then set the feature set to "1" else set to "0".

**URL Redirection:**
The presence of "//" in the URL path means that the user will be redirected to another webpage. If "//" presents in the URL path then feature is set to "1" else set to "0".

**Number of Dots in Hostname:**
Average no. of dots in legitimate websites is 3. If the dots are more than 3 then feature set to "1" else set to "0".

**HTTPS tokens in the URL:**
If the HTTPS token present in the URL then feature is set to "1" else set to "0".

**Prefix or Suffix separated by (-) to domain:**
The use of dash (-) symbol in legitimate URLs is rare. Attackers uses dash (-) symbol to domain name so that any users can feel that they are dealing with the legitimate URL. If the dash (-) symbol is present in the domain name then feature is set to "1" else set to "0".

**URL Shortening services (Tiny URLs):**
Tiny urls service allows attackers to hide long Phishing URL by making it short. The aim is to redirect the user to phishing websites. If the URL is small using shortening services (like bit.ly) then Feature is set to "1" else set to "0".

**Length of Hostname:**
It has been found that the length of legitimate site to be 25. If the length of the URL is more than 25 then feature is set to "1" else set to "0'.

**Website Rank:**
We have analyzed and compared the rank of the websites with the First 1 Lakh website of Alexa Database. If rank of the site is greater than 1 Lakh then feature is set to "1" else set to "0".

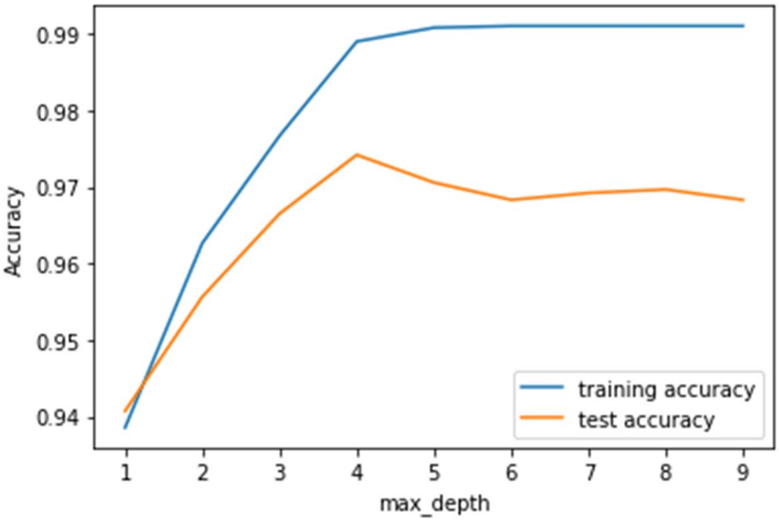
**Many more features are extracted just like the above and the machine learning algorithms are trained to detect the phishing URLs.**

## III. IMPLEMENTATION AND RESULT

**List Of Machine Learning Algorithms**

- Decision Tree
- Random Forest
- K-Nearest Neighbors
- Logistic Regression
- Naïve Bayes Classifier
- Support Vector Machine
- XGBoost Classifier
- Multi-Layer Perceptron
- CatBoost Classifier
- Gradient Boosting Classifier

Above are the algorithms that is trained based on the various features by setting the feature as "1" or "0". Each algorithm performs and results the output of its performance.



Figures 1.1

The figure1.1 shows the Training and Testing accuracy of Gradient Boosting Classifier.

Checking the Feature importance in the model (Figure 1.2)

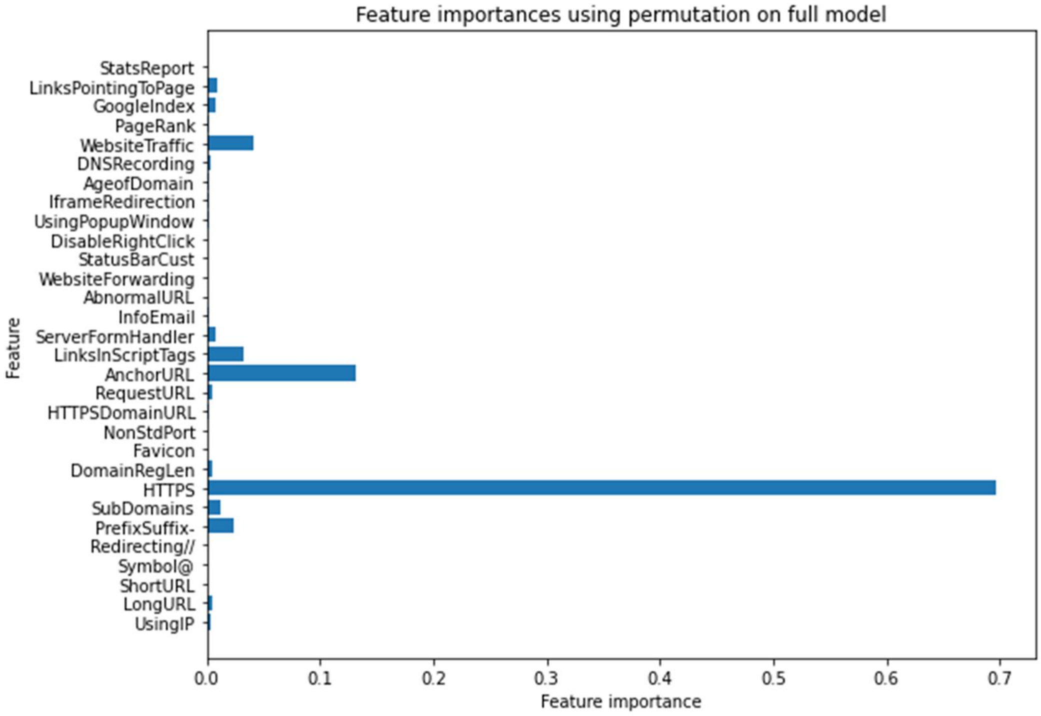
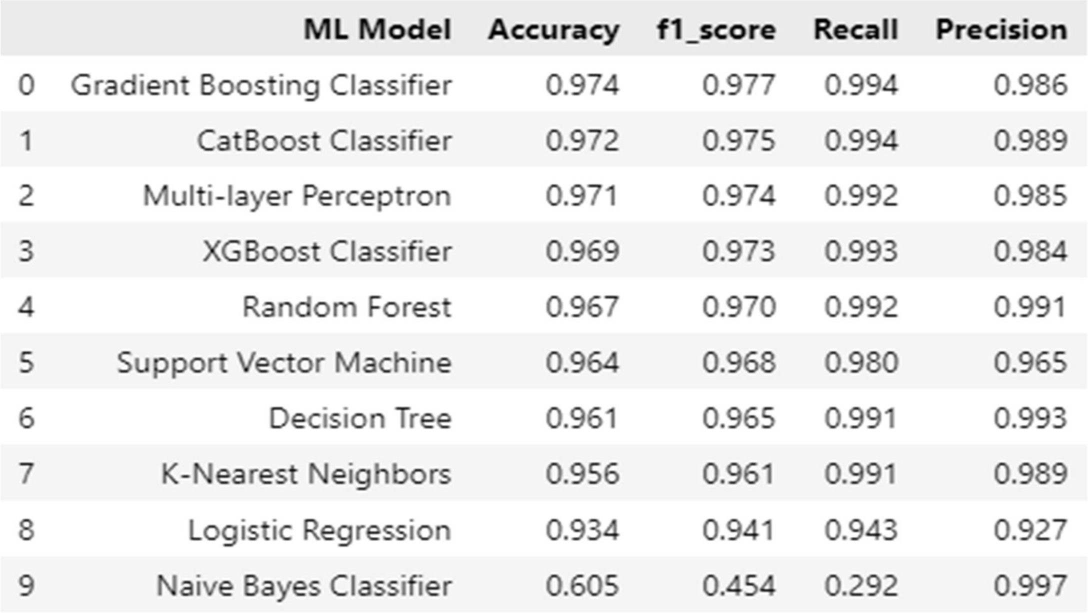


Figure 1.2

DOI: 10.48175/IJARSCT-15248

## IV. RESULT



| | ML Model | Accuracy | f1_score | Recall | Precision |
|---|---|---|---|---|---|
| 0 | Gradient Boosting Classifier | 0.974 | 0.977 | 0.994 | 0.986 |
| 1 | CatBoost Classifier | 0.972 | 0.975 | 0.994 | 0.989 |
| 2 | Multi-layer Perceptron | 0.971 | 0.974 | 0.992 | 0.985 |
| 3 | XGBoost Classifier | 0.969 | 0.973 | 0.993 | 0.984 |
| 4 | Random Forest | 0.967 | 0.970 | 0.992 | 0.991 |
| 5 | Support Vector Machine | 0.964 | 0.968 | 0.980 | 0.965 |
| 6 | Decision Tree | 0.961 | 0.965 | 0.991 | 0.993 |
| 7 | K-Nearest Neighbors | 0.956 | 0.961 | 0.991 | 0.989 |
| 8 | Logistic Regression | 0.934 | 0.941 | 0.943 | 0.927 |
| 9 | Naive Bayes Classifier | 0.605 | 0.454 | 0.292 | 0.997 |

The above figure shows the result of all the algorithms used in the detection and the accuracy of the model is between 0.605 to 0.974.
As we can see that the Gradient Boosting Classifier gives the highest accuracy = 0.974, f1 score = 0.977, Recall = 0.994, Precision = 0.986.

**Advantages**

- This system can be used by many E-commerce or other websites in order to have a good customer relationship
- Users can make the transaction securely
- Machine learning algorithms used in this system provides better performance compared to other traditional classification algorithms.
- Ensures a secure online experience for users.
- With the help of such system anyone can purchase the products online without any hesitation.

## V. CONCLUSION

Day by day internet users are increasing and they can be at serious risk from phishing. Because of the rapid development and spreading the phishing methods web security faces a lot of issues and it has been difficult to identify phishing URLs. This paper aims to enhance the method of identifying fraudulent URLs to detect the phishing websites using machine learning technology. Some of the feature of phishing dataset like “HTTPS”, “AnchorURL”, “WebTraffic”, have more importance o classify the URL is phishing or not. We have achieved 0.974 which is the highest accuracy using Gradient Boosting Classifier and hence reduces the chances of malicious attachment. And also, the result shows that more training data provided gives better performance.

## REFERENCES

**[1].** https://www.ijcaonline.org/archives/volume181/number23/mahajan-2018-ijca-918026

**[2].** https://www.sciencedirect.com/science/article/abs/pii/S0965997822001892#:~:text=To%20create%20a%20machine%20learning,model%20is%20the%20next%20step.

**[3].** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8504731/

DOI: 10.48175/IJARSCT-15248