**Phishing URL Prediction Using AI-ML**

**Department of Instrumentation and Control Engineering**

**Digvijay Patil1, Dr. Prof. Shilpa Sondkar2**

1Student, BTech Vishwakarma Institute of Technology, Pune

2Head of Instrumentation Department, Vishwakarma Institute of Technology, Pune

**Abstract:** - Phishing attack is the simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password, and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analysing various features of legitimate and phishing URLs. Through machine learning algorithms one can improve the accuracy of the prediction. We identified features that phishing site URLs contain. The proposed method employs those features for phishing detection. We shall talk about various machine learning algorithms which can help in decision making and prediction. We shall use more than one algorithm to get better accuracy of prediction.

**Keywords*:*** Phishing Website, Logistic regression, Data classification, Naive bayes, Machine Learning, legitimate, URL, feature extraction***.***

1. **INTRODUCTION**

Nowadays Phishing has become a main area of concern for security researchers because it is not difficult to create fake websites which look so close to legitimate websites. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attacks. Main aim of the attacker is to steal bank account credentials. In United States businesses, there is a loss of US$2billion per year because their clients’ become victims of phishing. In the 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $5 billion.

Phishing imitates the characteristics and features of emails and makes it look the same as the original one. It appears similar to that of the legitimate source. The user thinks that this email has come from a genuine company or an organization. This makes the user to forcefully visit the phishing website through the links given in the phishing email. These phishing websites are made to mock the appearance of an original organization website. The phishers force users to fill up the personal information by giving alarming messages or validate account messages exactly so that they fill up the required information which can be used by them to misuse it. They make the situation such that the user is not left with any other option but to visit their spoofed website.

1. **RELATED WORK**

Many researchers have previously been carried out in this field of phishing detection. We have gathered the information from various such works and have profoundly reviewed them which has helped us in motivating our own methodologies in the process of making a more secure and accurate system.

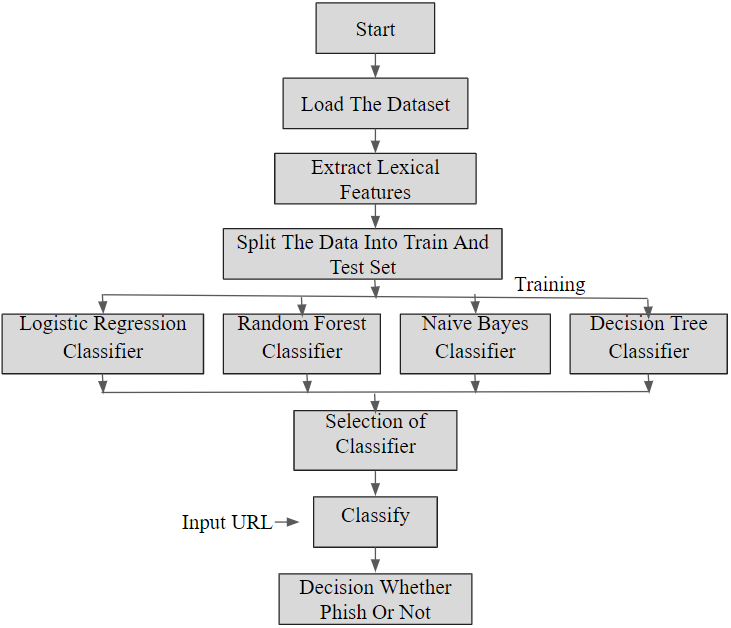
[1] the authors proposed a system with a collection or set of Hybrid features to classify websites based on machine learning algorithms. The main feature set is extracted using the cumulative distribution gradient technique, while the data perturbation ensemble technique is used to extract the secondary feature set. The algorithm used for training the classifier is Random Forest in association with ensemble learner identifies the phishing websites with a precision of 94.6 percent.

[2] detected phishing websites using a machine learning algorithm. The feature set included six features based on the website structure and was chosen after a comparative study by the authors. The classifier was trained using Support Vector Machine which worked effectively to classify websites whether legitimate or phishing. The model presented obtained an accuracy of 84 percent for the classification of websites.

[3] The authors designed a browser extension to detect phishing websites. The system used multiple machine learning algorithms which included Random Forest, Support Vector Machine (SVM), and k-Nearest Neighbour (KNN) to train the classifier to achieve higher precision by doing a comparative study. The feature set included a content-based approach for extracting the JavaScript and HTML features of the websites. The dataset was imported from UCI-Machine Learning Repository and boasted a 22-feature classification technique to detect phishing websites.

[4] A phishing website detection method using reduced feature classification. The extracted features were analysed using Support Vector Machine (SVM) and Logistic Regression algorithms. Out of the total 30 features identified, 19 features were selected and used for classification. The model was implemented using Big Data and the Dataset was obtained from the UCI Irvine machine learning repository. Between the two algorithms used, Support Vector Machine (SVM) showed better performance and accuracy of 95.62%.

1. **METHODOLOGY**

****

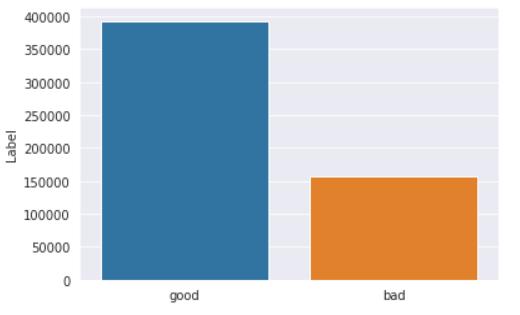
**Figure 1: Flow chart of the proposed system**

This section explains the methodology to detect Phishing Websites using ML, and also explains the proposed framework. The experiment was carried out by using ML approaches. ML approaches can be applied in two ways. The first one is supervised learning, and another is unsupervised learning.

**A. Dataset Description**

The dataset is one of the most critical parts of our study. A dataset is nothing but the table containing information about phishing and legitimate websites—the dataset for our proposed model obtained from Kaggle. Kaggle is one of the most popular public repositories with a tremendous amount of dataset collection which can be used for training machine learning models. In that dataset containing 5,49,346 entries and two columns. First column contains the URL, and the second column is the label column. In the label column is the prediction column which has two categories. First one is a good URL which means the URL does not contain malicious information and second is a bad URL.

**Distribution of Good and Bad URL**: - For different features we put different rules based on the analysis of good and bad websites. Data visualization is an easier way of presenting the data, however complex it is, to analyse trends and relationships amongst variables with the help of pictorial representation.



**Figure 2: Ratio of Good and Bad URL**

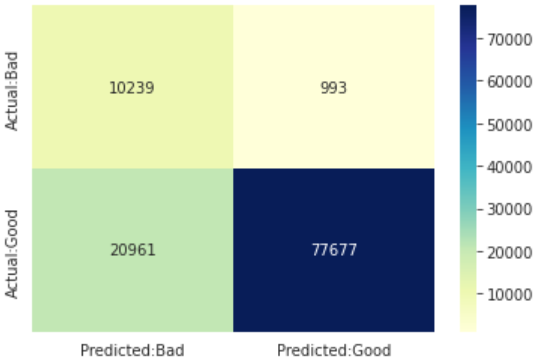
**B. Data pre-processing**: -

Feature scaling is the process of normalizing or standardizing the independent variables of the training dataset to a fixed range, to handle variance in the values among different independent variables. Splitting the dataset into two portions, one for training and one for testing is very important. It is vital to train a model with a subset of the full dataset and test model with the rest to evaluate the model performance satisfactorily. We split the dataset into 80:20 ratio as follows: 80% of the dataset used for training and 20% of dataset for testing using a stratified sampling technique. We did the train test split using the Scikit-Learn library in the Python programming language.

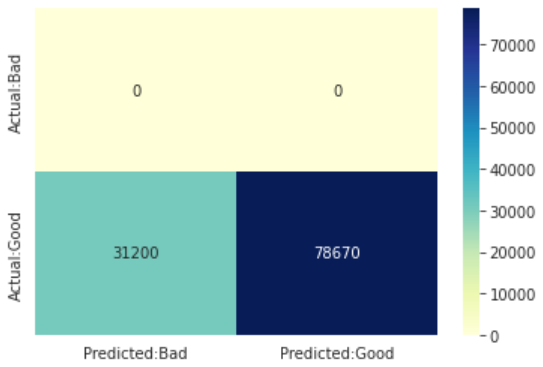
**C. Apply Machine Learning-**

When data has been ready, we apply Machine Learning Technique. Four machine learning classification models Decision Tree, random forest, naive bayes and Logistic Regression Have been selected to detect phishing websites.

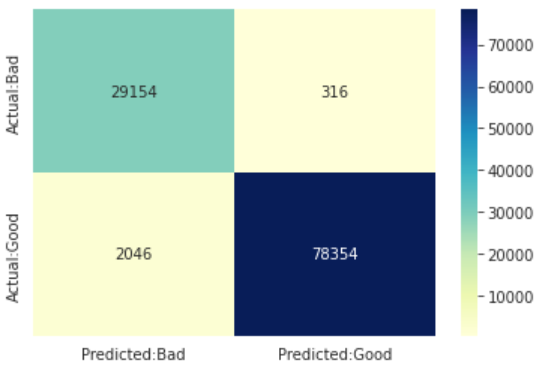
**1.Decision Tree Algorithm**: - Decision tree algorithm is easy to understand and also easy to implement. Decision tree begins its work by choosing the best splitter from the available attributes for classification which is considered as a root of the tree. Algorithm continues to build the tree until it finds the leaf node. Decision tree creates a training model which is used to predict target value or class in tree representation each internal node of the tree belongs to attribute and each leaf node of the tree belongs to class label. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data).In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record’s attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

**Figure 3: Confusion Matrix For DT**

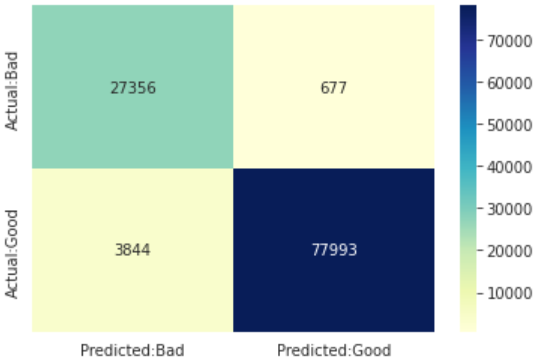
**2.Random Forest Algorithm: -** Random Forest algorithm creates the forest with a number of decision trees. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems. High number of trees gives high detection accuracy. Creation of trees is based on the bootstrap method. In bootstrap method features and samples of dataset are randomly selected with replacement to construct a single tree. Among randomly selected features, the random forest algorithm will choose the best splitter for the classification and like the decision tree algorithm; Random Forest algorithm also uses Gini index and information gain methods to find the best splitter. This process will continue until a random forest creates a number of trees. Each tree in the forest predicts the target value and then the algorithm will calculate the votes for each predicted target. Finally, the random forest algorithm considers a highly voted predicted target as a final prediction.

**Figure 4: Confusion Matrix For RF**

**3. Naive Bayes Algorithm: -** Naive Bayes is a simple probabilistic classifier based on conditional probability. This classifier calculates the posterior probability for each class and assigns the sample to the class with the maximum probability and also uses basic Bayesian theorem and propagates a firm independence that is present between features. Occurrence of features along with their inter-relations is calculated in corpus consideration. It uses one class label. In NB, correlation of all neglected attributes is considered to be independent. Naïve Bayes algorithms are often used in sentiment analysis, spam filtering, recommendation systems, etc. They are quick and easy to implement but their biggest disadvantage is that the requirement of predictors to be independent.

**Figure 5: Confusion Matrix For NB**

**4. Logistic Regression Algorithm: -** Logistic is a statistical model. It uses a logistic function to model a binary dependent variable. In our regression analysis, we use 80% labelled data to acquire a logistic function to predict whether a website is a legitimate or a phishing website. Logistic regression performs better when the data is linearly separable. It does not require too many computational resources as it’s highly interpretable. There is no problem scaling the input features—It does not require tuning. It is easy to implement and train a model using logistic regression. It gives a measure of how relevant a predictor (coefficient size) is, and its direction of association (positive or negative).

**Figure 6: Confusion Matrix For LR**

1. **MODEL BUILDING**

This is the most important phase which includes model building for prediction of phishing websites. In this we have implemented various machine learning algorithms which are discussed above for phishing website prediction. Procedure of Proposed Methodology

**Step1:** Import required libraries, Import phishing website dataset.

**Step2:** Pre-process data to remove missing data.

**Step3**: Perform a percentage split of 80% to divide the dataset as Training set and 20% to Test set.

**Step4:** Select the machine learning algorithm.

**Step5:** Build the classifier model for the mentioned machine learning algorithm based on the training set.

**Step6:** Test the Classifier model for the mentioned machine learning algorithm based on the test set.

1. **RESULTS AND ANALYSIS**

|  |  |
| --- | --- |
| **Algorithm** | **Accuracy** |
| Random Forest | 94.67 |
| Logistic Regression | 96.58 |
| Support Vector Machine Classifier | 95.23 |
| Extreme Gradient Boost | 94.97 |

**Table 1: Accuracy Score of Classifiers**

1. **CONCLUSION AND FUTURE SCOPE**

* It will help us to detect phishing sites and phishing attacks.
* It can avoid financial loses of people due to phishing as well as spyware.
* It will save some amount of money and time of many users.
* It decreases the work of cybercrime police cops.
* The user must enter link of any site then it will be able to easily predict its type.

1. **REFERENCES**
2. Kang Leng Chiew, Choon Lin Tan, KokSheik Wong, Kelvin SC Yong, and Wei King Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," Information Sciences,2019.
3. Pan, Ying, and Xuhua Ding. ―Anomaly based web phishing page detection. ‖ In Computer Security Applications Conference, 2018. ACSAC'18. 22nd Annual, pp. 381-392. IEEE, 2018.
4. Desai, J. Jatakia, R. Naik, and N. Raul, “Malicious web content detection using machine learning,” RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc.