Reference number:46S_BE_1406

Phishing Website Detection Plugin Using Machine Learning

SDM Institute of Technology Ujire - Computer Science and Engineering

Student names:

Ms.Anushree SM

Ms.Anvitha Ramesh

Ms.Ashwini C Bhat

Ms.Disha Shetty

Guid name:

Prof.Pradeep G S

1. Keywords

Ensemble Model, Stacking, Machine Learning, Phishing, Cyber Security, Data Breach, Feature Engineering and Blacklist.

2. Introduction

Phishing is a type of cyber-security attack where malicious actors send messages pretending to be a trusted entity. Phishing messages manipulate a user, causing them to perform actions like installing or clicking a malicious file or revealing sensitive information. Phishing attackers use a variety of techniques such as link manipulation, filter evasion, website forgery & covert redirect. Phishing attacks have become a significant concern owing to an increase in their numbers. A typical phishing attack technique involves using a phishing website, where the attacker lures users to access fake websites by imitating the names and appearances of legitimate websites, such as eBay, Facebook, and Amazon. It is difficult for the average person to distinguish phishing websites from normal websites because phishing websites appear similar to the websites they imitate. In many cases, users do not check the entire website URL, and, once they visit a phishing website, the attacker can access sensitive and personal

information. With the growth in the field of e-commerce, phishing attack and cyber crimes are rapidly growing. Attackers use websites, emails, and malware to conduct phishing attacks

3. Objectives

The objectives of the proposed project are as follows:

1. To train the ML Classifier models like Random Forest, Cat Boost, Light GBM and their ensemble model using the datasets of phishy & legitimate websites..
2. To run comparative analysis by calculating accuracy of each training model to know the best suitable model for phishing website detection.
3. To create a phishing detection system using the best suitable model.
4. To develop an interactive web browser plug-in for the real-time detection & blocking of Phishing website.

4. Methodology

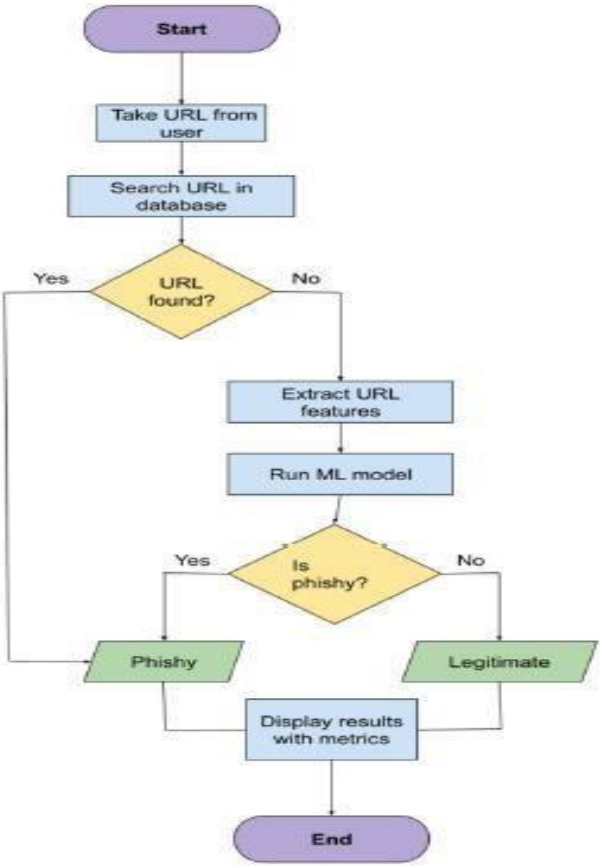

Figure 4: Flowchart of the proposed system.

A. Method

The proposed phishing detection system is implemented using the following steps:

1) **Data Acquisition & Preprocessing:** This step involves collection of datasets of phishy and legitimate websites from open source platforms like Kaggle and then preprocessing of data.

2) **Database Creation:** This process follows blacklist methodology of phishing detection in which database containing blacklist of illegal websites is created. Any phishy website detected by our model will be added to blacklist database so that when the user encounters same website, the site is immediately blocked with faster response time.

3) **Training of URL classifier model:** This involves selecting various classifier algorithms and training them using URL Features, running a comparative analysis of performance and choosing best suited model. We test various algorithms like Random Forest, CatBoost, Light GBM and their ensemble model, train and analyze the performance. Finally the best suited model which is the ensemble model is chosen.

4) **Developing Website content analysis module:** Even after classification of website as legitimate, there is a possibility that a legitimate website may be housing unwanted images, icons and pop-ups. Hence a website content analysis is which provides user with information regarding presence of suspicious elements.

5) **Extraction of URL Features & Output Prediction:** The user provides URL of the required website as input. Relevant features are extracted from this URL and a dataframe is created. The ensemble model takes the extracted features of the given website to predict whether the URL is suspicious or not. If the URL is found to be suspicious then the user is warned and the blacklist database will be updated automatically.

B. Model

Model for detecting the phishing website is developed using Ensemble stacking of these 3

algorithms.

1. Random Forest

A supervised learning technique which builds decision trees by taking different samples as input and considers the major number of votes for classification and in case of regression it considers the average. A accurate result is obtained if the forest

contains a large number of trees and the problem of over fitting can be tackled. Random forest works by creating a
decision tree in the beginning and then makes predictions for every single tree created.

2. Cat Boost

Cat boost is a gradient boosting algorithm that uses a decision tree as its base estimator and
combines multiple weak decision trees to create a strong ensemble model. It can handle categorical data by using an encoding algorithm called ordered boosting, which assigns a numerical value to each category based on the target variable. Cat boost also uses a unique method for handling missing by using a default prediction value that is updated as the algorithm progresses. Additionally, it incorporates various techniques to improve prediction accuracy, such as a novel method for calculating feature importance and a form of regularization called gradient-based one-side sampling (GOSS) to prevent overfitting. Catboost is a popular machine learning tool for various applications due to its effectiveness in handling categorical data and handling missing values.

3. Light GBM

Light GBM is a gradient boosting framework that uses tree based learning algorithm. Light GBM grows tree vertically while other algorithm grows trees horizontally meaning that Light GBM grows tree leaf-wise while other algorithm grows level-wise. It will choose the leaf with max delta loss to grow. When growing the same leaf, Leaf-wise algorithm can reduce more loss than a level wise algorithm.

5. Result and Conclusion

Bar graph representing accuracy score of Algorithms.

Figure 5 represents the accuracy score of all 3 algorithms i.e,: Random Forest, Cat Boost, Light GBM and their Ensemble Model respectively. The highest accuracy score among the considered 4 models is by Ensemble Model that is 95.14%.

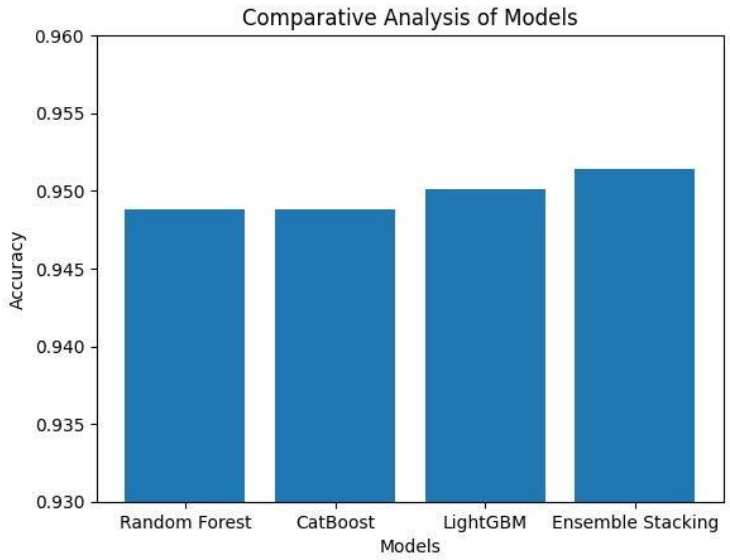

Figure 5:Bar graph representing accuracy score

In this project, we have used a dataset containing sufficiently large amount of data about every possible URL structure.The attributes considered here were like Web Traffic, Domain Age, Google and many more. These attributes were used and classified using 4 models. Among the models, Ensemble model proved to be much more efficient with an accuracy of around 95.14%. The paper proposes an efficient method for improvising current phishing detection techniques by combining both blacklist approach and machine learning approach. The use of blacklist approach reduces the response the since the URL will be searched in the database before going through feature extraction and being passed to the classifier for prediction. This application is aimed at providing both higher accuracy and speed so that anyone who uses “Citadel” the phishing detection plugin finds its efficiency in ease of use. It can be used by every user to effectively shield from malicious attackers. The simple design helps even the novice user to understand the usage and easily access the application.

6. Scope of Future Work

The project can be further improvised by developing a desktop application which could be downloaded to user’s system. This can help user get relevant messages and warnings and also changes in any features could be notified immediately. The project can also be incorporated with an API service for developers to use the model for requesting result of determining the legitimacy of websites in their applications.