

PHISHING DOMAIN DETECTION

(Machine Learning)

# ARCHITECTURE DOCUMENT

Sunil Bharadhwaj C

November 2023

**INTRODUCTION:**

Phishing Domain Detection is a technique that enables us to predict whether a domain is real or fake.

## PROBLEM STATEMENT:

Phishing is a type of fraud in which an attacker impersonates a reputable company or person through email or other communication channels to obtain sensitive information such as login credentials or account details. Attackers prefer phishing because convincing someone to click a seemingly authentic malicious link is easier than bypassing a computer's security measures. The primary objective is to predict the authenticity of domains, distinguishing between genuine and malicious ones.

## APPROACH:

In the project, we've completed the classical machine learning tasks, including Data Exploration, Data Cleaning, Feature Engineering, Model Building, and Model Testing. We experimented with various machine learning algorithms, such as Logistic Regression, Decision Tree Regressor, and Random Forest. After thorough evaluation, we determined that XGBoost is the best-fit model for the project.

## DATASET:

Dataset Information:

The dataset consists of both legitimate and phishing website instances. Each website is represented by a set of features that help determine whether a website is legitimate or not. This data serves as input for machine learning processes.

* Full Variant Dataset - dataset\_full.csv:
* Total number of instances: 88,647
* Number of legitimate website instances (labeled as 0): 58,000
* Number of phishing website instances (labeled as 1): 30,647
* Total number of features: 111
* Small Variant Dataset - dataset\_small.csv:
* Total number of instances: 58,645
* Number of legitimate website instances (labeled as 0): 27,998
* Number of phishing website instances (labeled as 1): 30,647
* Total number of features: 111
* This data provides a valuable resource for machine learning and predictive modeling tasks related to website legitimacy and phishing detection.

## WORK FLOW:

## Data Collection: The project starts by collecting a dataset of URLs, encompassing both legitimate and potentially malicious websites. This dataset serves as the foundation for training and testing the predictive model.

## Feature Extraction:

## Character Length and Vowels: For each URL in the dataset, the project extracts the length of the URL and counts the number of vowels, which can serve as simple indicators of suspicious characteristics.

## Sender Policy Frameworks (SPF): SPF records associated with the domains are retrieved to check if the domain's email sender policy aligns with known legitimate standards.

## Website Creation and Expiration Dates: The project extracts information about the domain's creation and expiration dates to assess the age and potential longevity of the website.

## Servers and Redirects: Server information is gathered to identify unusual or suspicious server configurations. Additionally, redirects are tracked to uncover potential hiding or obfuscation attempts.

## Response Time: The response time of each website is measured to assess its performance and responsiveness.

## The extracted features, including URL characteristics and data about SPF records, server details, and response times, are used to create a dataset.

## An XGBoost classification model is trained on this dataset. XGBoost is chosen for its ability to handle complex, high-dimensional data and its capacity to provide robust predictions.

## Phishing Detection:

## Once the XGBoost model is trained, it can predict whether a given URL is a legitimate or phishing domain based on the extracted features.

## The system generates predictions for each URL in real-time, classifying them as safe or potentially harmful.

## Benefits:

## Enhanced Security: Phishing Domain Detection using XGBoost provides an added layer of security for users, helping to identify and avoid malicious websites.

## Automation: The system can automatically assess URLs, making it valuable for email filtering, web browsers, and other internet security applications.

## Real-time Protection: By measuring response times and analyzing URL characteristics, the system can provide real-time protection against emerging phishing threat.

## CONCLUSION:

## This project harnesses data-driven techniques, feature extraction, and machine learning to create a robust tool for detecting and guarding against phishing domains. The XGBoost model, trained on a diverse dataset, holds the potential to offer real-time, dependable protection against malicious websites, thereby enhancing the safety of online experiences for users.

## The results from phishing domain detection, utilizing the random forest model, are highly promising, showcasing an accuracy rate of 96.7%, precision of 95%. These metrics reflect the model's outstanding performance in accurately identifying phishing domains. They affirm the model's high accuracy in recognizing malicious domains and its reliability in distinguishing whether a domain is involved in phishing activities or not.

## Overall, this model serves as a valuable tool for shielding users against phishing threats. Its exceptional accuracy and precision make it a compelling choice for deployment in real-world scenarios