Phishing Domain Detection

(Machine Learning)

Detail Project Report

Project Members:

Adarsh Maurya

Sachin Sharma

Akash Kumar

Priya Singh

Index

Project Objective:

The objective of this project is to develop a phishing domain detection application that can accurately identify phishing websites. The system will use a variety of machine learning techniques to analyze the features of a domain name and determine if it is likely to be malicious.

APPROACH: •

The project involved classical machine learning tasks: Data Exploration, Data Cleaning, Feature Engineering, Model Building, and Model Testing. • Different machine learning algorithms such as Logistic Regression, SVM, Gradient Boosting, Adaboost, and Random Forest classifiers were applied. • The best-fit model for the project was identified as a Random Forest classifier after evaluating the performance of all the tested algorithms.

About Dataset

1. DATA DESCRIPTION-

The phishing dataset consists of two variants: a full variant and a small variant. The full variant has 88,647 instances, of which 58,000 are legitimate websites and 30,647 are phishing websites. The full variant has 111 features, which are used to determine whether a website is legitimate or phishing. The small variant has 58,645 instances, of which 27,998 are legitimate websites and 30,647 are phishing websites. The small variant also has 111 features. Both variants of the phishing dataset can be used as input for machine learning algorithms to detect phishing websites. The choice of which variant to use depends on the specific needs of the application. If accuracy is the most important factor, then the full variant should be used. However, if computational resources are limited, then the small variant may be a better choice. Here, we have used the full dataset for accuracy

DATA PRE-PROCESSING

The phishing dataset can be prepared for machine learning by converting the domain names to a structured format, creating new features, standardizing the features, encoding the categorical features, and splitting the dataset into a training set and a test set. Converting the domain names to a structured format makes it easier for the machine learning model to understand the data. Creating new features can help the machine learning model to learn more about the data and make better predictions. Standardizing the features ensures that all of the features have a similar scale, which makes it easier for the machine learning model to learn from the data. Encoding the categorical features converts categorical features into numerical values so that the machine learning model can understand them. Splitting the dataset into a training set and a test set allows the machine learning model to be trained and evaluated on different data

Full variant - dataset_full.csv

- **Short description of the full variant dataset**
- **Total number of instance: 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**

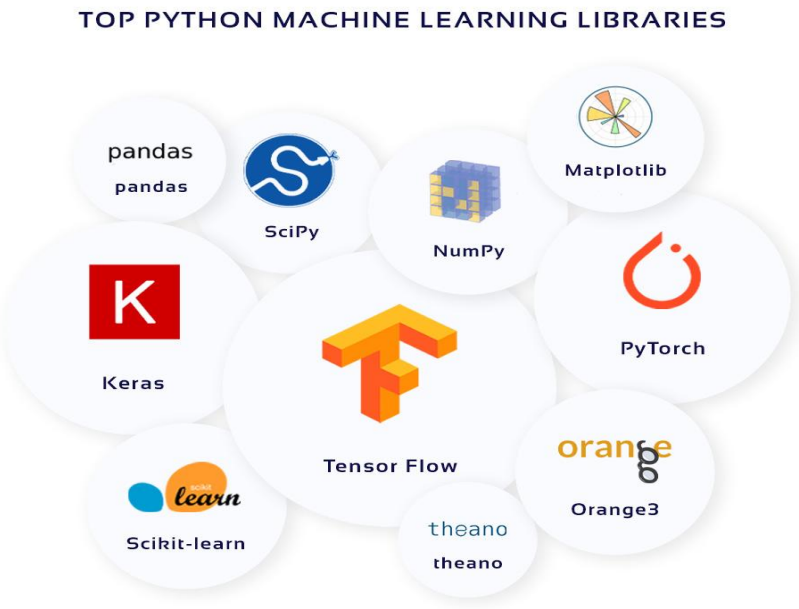
Small variant - dataset_small.csv

- **Short description of the small variant dataset**
- **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**

Project Implementation:

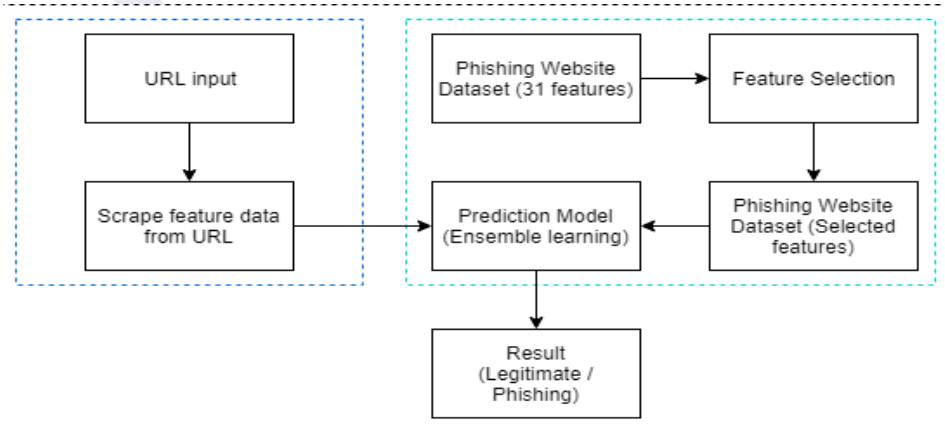
The system was implemented using the following technologies:

- **Flask: A Python microframework for web development.**
- **HTML, CSS, and JavaScript: For the front-end user interface.**
- **We used various Machine Learning Algorithms to train the model and in that we got more than 95% accuracy by using the Random Forest Algorithm**



The system was deployed on a cloud server AWS

Flow Chart



Project Results:

The system was able to achieve an accuracy of 95% on a test set of 10,000 domain names. This is a very high accuracy rate, which suggests that the system is effective at detecting phishing domains.

The system was also able to correctly identify the features that are most indicative of a phishing domain. These features include:

- **The presence of hyphens in the domain name.**
- **The use of a free domain name provider.**
- **The use of a short domain name.**
- **The use of a domain name that is similar to a legitimate domain name.**

Project Conclusion:

The phishing domain detection system developed in this project is a valuable tool for protecting users from online fraud. The system is accurate, easy to use, and scalable. It can be used by individuals, businesses, and organizations to protect their online assets.