PHISHING PREDICTION

Predict phishing website by ML

ABSTRACT

The basic idea of this ML model to provide a safe browsing environment for the IT industry employees and other people also, all industries are moving to online for scaling which lead to increase in websites as well as the phishing websites.

REVISION NUMBER – 1.0

Authored By:
Nihar Ranjan Samal
M.Sc. Data Science

Content

Document Version Control

Date	Version	Description	Author
21/09/2022	1.0	Abstract, Introduction, General Description, Design Flow	Nihar Ranjan Samal

Abstract

The basic idea of this ML model to provide a safe browsing environment for the IT industry employees and other people also, all industries are moving to online for scaling which lead to increase in websites as well as the phishing websites. Phishing is popular among attackers because it’s easier to trick (social engineering) someone to click on a link. It is very important to know if a website is fake or legitimate, mistake from employee may lead to a saviour lose to the company.

1.Introduction

1.1 Why this HLD Document?

The main purpose of this HLD document is to feature the required details of the project and supply the outline of the Model Creation, Evaluation and Deployment. This additionally provides the careful description on however the complete project has been designed end-to-end. The HLD will:

- Present of the design aspects and define them in detail.
- Describe the user interface being implemented.
- Describe the hardware and software interfaces.
- Describe the performance requirements.
- Include design features and architectural design of the project.
- List and describe the non - functional attributes like:
    - Security
    - Reliability
    - Maintainability
    - Portability
    - Reusability
    - Resource Utilization

1.2. Scope

The HLD documentation presents the structure of the system, such as database design, architectural design, application flow and technology architecture. The HLD uses non-technical terms to technical terms that can be understandable to the administrator of the system. The basic idea is to detect the website is fishing website or not.

1.3. Definition

Term	Description
Dataset	Collected information for prediction
Jupyter-Notebook	It is an interactive computational environment, in which you can combine code execution, rich text, mathematics, plots and rich media.
AWS	AWS is cloud platform that enables developers to build, run, and operate applications.

2. General Description

2.1. General Perspective

The phishing site prediction may be a machine learning model that helps the user find if a website is a fraud website or a healthy website and help them to not to visit that website.

2.2. Problem Statement

Phishing is a type of fraud in which an attacker impersonates a reputable company or person to get sensitive information such as login credentials or account information via email or other communication channels. Phishing is popular among attackers because it is easier to persuade someone to click a malicious link that is authentic than it is to break through a computer's protection measures.

The main goal is to predict whether the domains are real or malicious.

2.3. Proposed Solution

To solve the problem, we have created a User interface for taking the input from the user to predict the Phishing Website using our trained ML model after processing the input and at last the output (predicted value) from the model is communicated to the User.

2.4. Technical Requirements

As technical requirements, we don’t need any specialized hardware for virtualization of the application. The user should have the device that has the access to the web and the fundamental understanding of providing the input.

2.5. Tools Used

- Python 3.9 is employed because the programming language and frameworks like NumPy, Pandas, Scikit – learn, *LightBGM* and alternative modules for building the model.
- Jupyter-Notebook is employed as IDE.
- For Data visualizations, seaborn and components of matplotlib are getting used.
- For information assortment prophetess info is getting used.
- Front end development is completed victimization HTML/CSS.
- Flask is employed for each information and backend readying
- GitHub is employed for version management.
- AWS is employed for deployment

2.6. Constrains

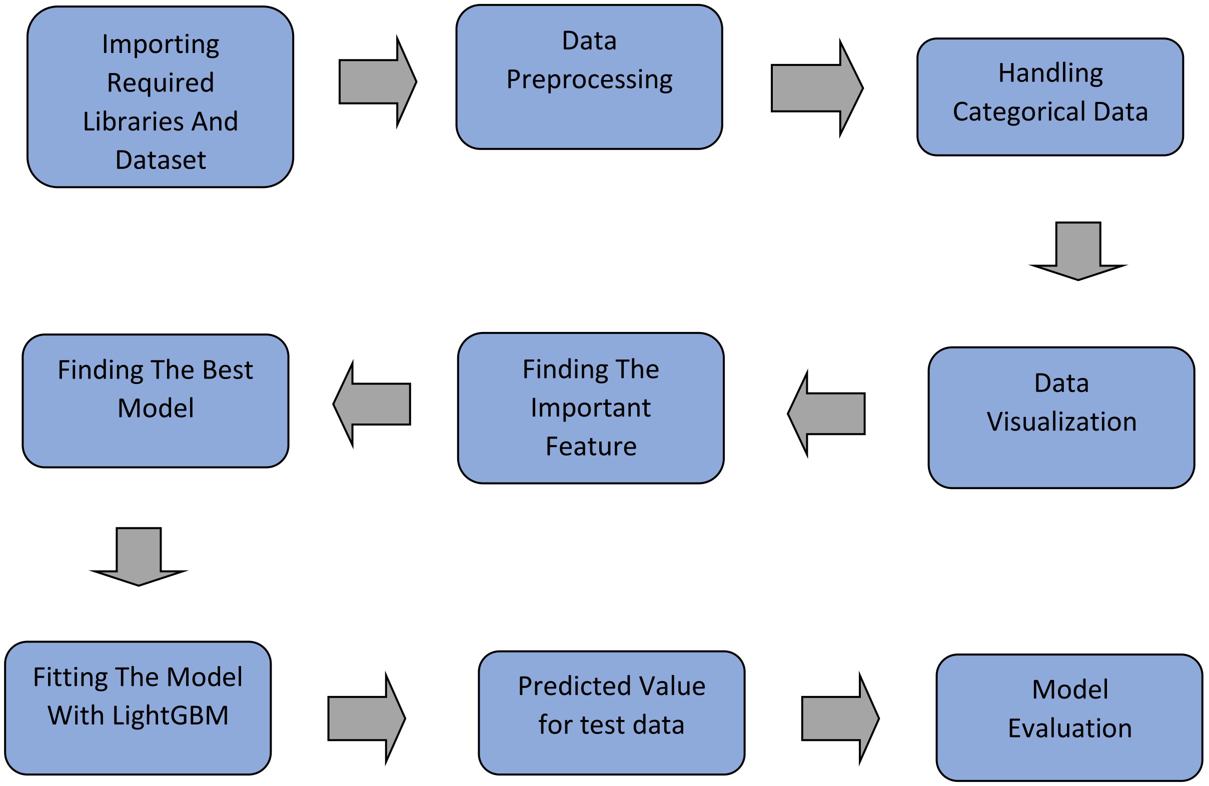
For multiple site prediction the user should upload csv file in the given format.

2.7. Assumption

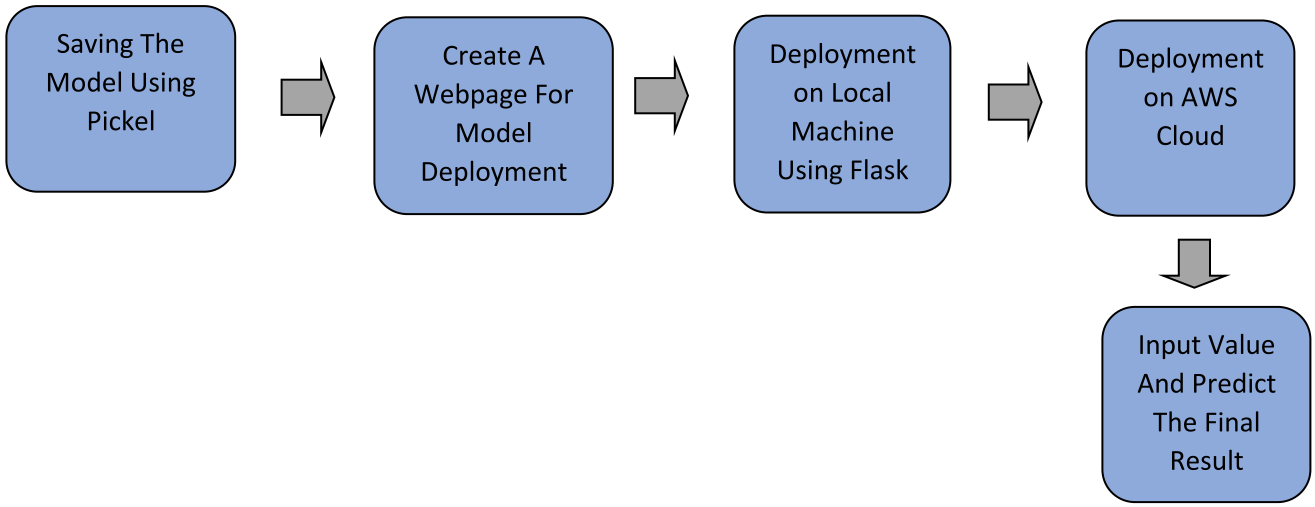
The main objective of the project is to implement the utility cases as for the new dataset that provides the user the ability to predict Phishing website. Machine learning model is employed for process the user input for prediction. It additionally assumed that each one aspects of this project have the flexibility to figure along within the approach the designer is expecting.

3. Design Flow

3.1. Modelling Creation and evaluation


Importing Required Libraries And Dataset
Data Preprocessing
Handling Categorical Data
Data Visualization
Finding The Important Feature
Finding The Best Model
Fitting The Model With LightGBM
Predicted Value for test data
Model Evaluation

3.2. Deployment Process


Saving The Model Using Pickel
Create A Webpage For Model Deployment
Deployment on Local Machine Using Flask
Deployment on AWS Cloud
Input Value And Predict The Final Result

3.3. Logging

In logging, at each if an error or an exception is occurred, the event is logged into the system log file with reason and timestamp. These helps the developer to debug the system bugs and rectifying the error.

3.4. Error Handling

Once the error is occurred, the reason is logged into the log file with timestamp to rectify and handle it.

4. Performance Evaluation

4.1 Reusability

The code written and the components used should have the ability to be reused with no problems.

4.2 Application Compatibility

The different parts of the system are communicating or using Python as an interface between them. All the components have its own tasks to perform and it is a job of a Python to ensure proper transfer of data.

4.3 Resource Utilization

When ant task is performed, it'll doubtless use all the process power offered till the process is finished.

4.4 Deployment

The model can be deployed using the any cloud services such as Microsoft Azure, Amazon web services, Heroku, Google cloud, etc.

Conclusion

The Phishing prediction will predict the website is fraud or not and ensuring the domain is right domain.