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 DPR Document?

The main purpose of this DPR documentation is to add the necessary details of the project and provide the description of the machine learning model and the written code. This also provides the detailed description on how the entire project has been designed end-to-end.

Key points :

- Describes the design flow
- Implementations
- Software requirements
- Architecture of the project
- Non-functional attributes like:
    - Reusability
    - Portability
    - Resource utilization

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.

3. 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.

3.1. 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

4. Data Requirements

The Data requirements is totally supported the matter statement and also the dataset is accessible on the Mendeley within the file format of (.csv).

4.1. Data Collection

The data for this project is collected from the Kaggle Dataset, the URL for the dataset is https://data.mendeley.com/datasets/72ptz43s9v/1

4.2. Data Description

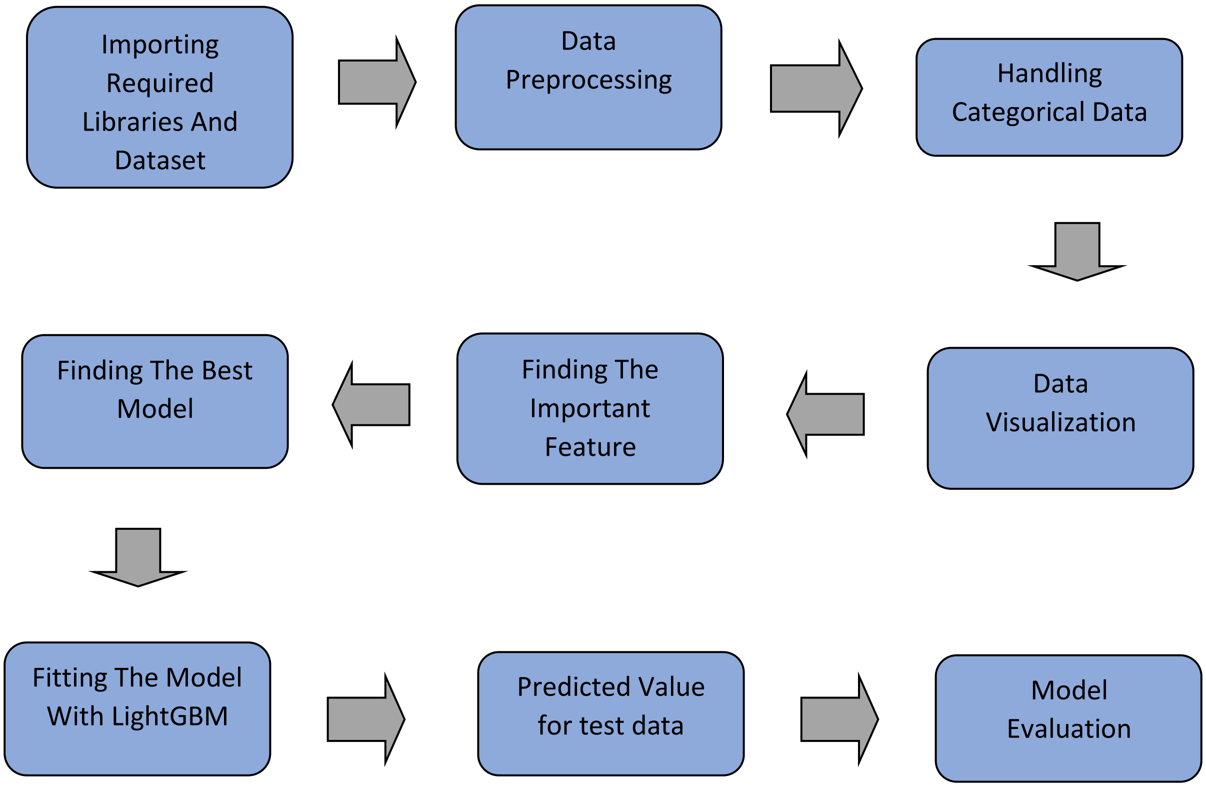
The dataset contains 112 variables. The dataset contains more than 8800 records and total size of the dataset is approximately 40 MB. There are 112 entries from qty_dot_url to phishing and of data type 1 float64 and 111 int64.

5. Data Preprocessing

- Checked for info of the Dataset, to verify the correct datatype of the Columns.
- Checked for Null values, because the null values can affect the accuracy of the model.
- Checking the distribution of the columns to interpret its importance.
- Now, the info is prepared to train a Machine Learning Model.

6. Design Flow

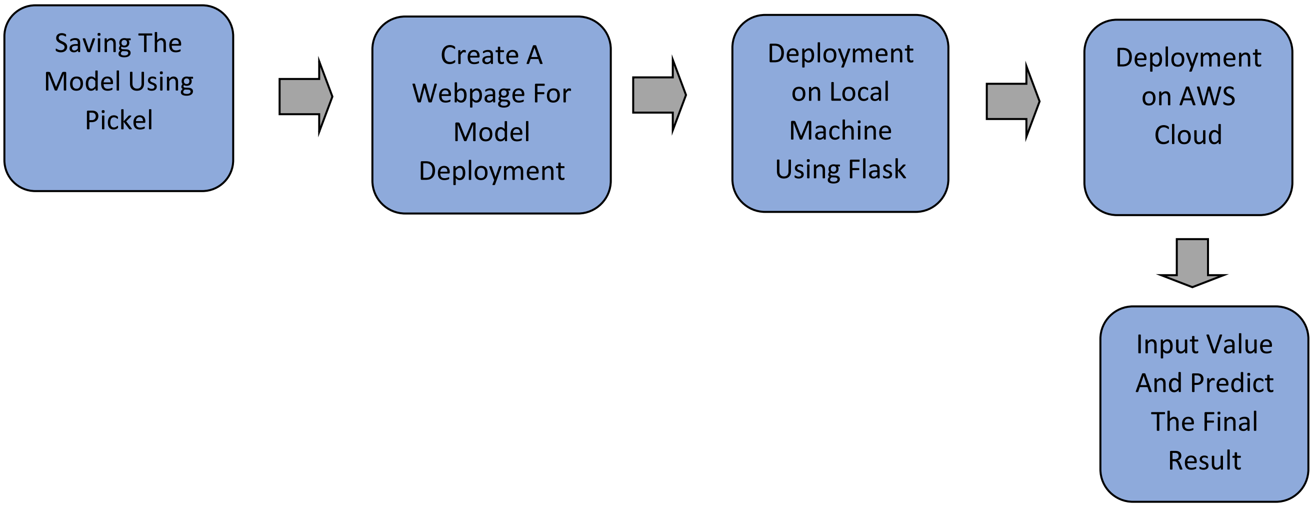
6.1. Modelling Creation and evaluation


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

6.2. UI Integration

Both CSS and HTML files are being created and are being integrated with the created machine learning model. All the required files are then integrated to the app.py file and tested locally.

6.3. 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

6.4. 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.

7. Data From User

The data from the user is retrieved from the created HTML web page.

8. Data Validation

The data provided by the user is then being processed by app.py file and validated. The validated data is then sent to the prepared model for the prediction.

9. Rendering The Result

The data sent for the prediction is then rendered to the web page.

10. Deployment

The tested model is then deployed to Heroku. So, users can access the project from any internet devices.

11. Conclusion

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

12. FAQs