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

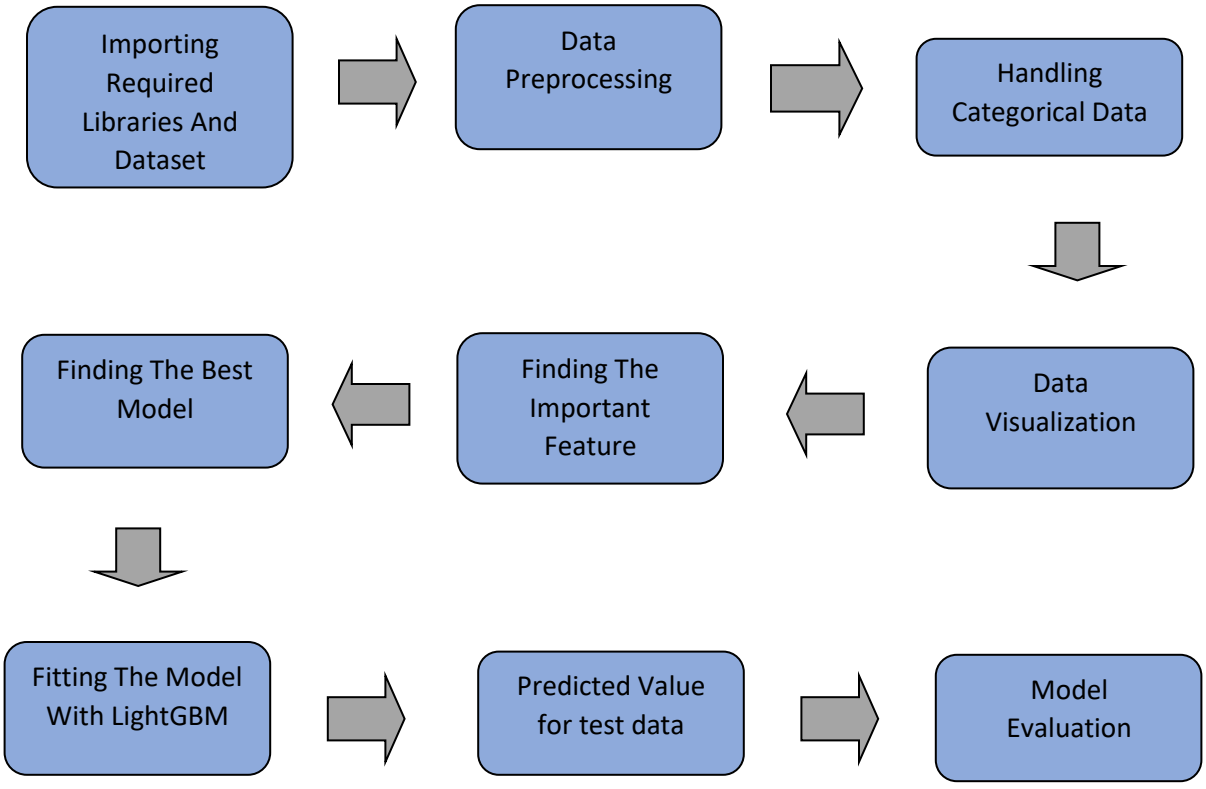
The basic idea of this ML model to provide a safe browsing environment for the IT industry employees and other people also, all industries are moving to online for scaling which lead to increase in websites as well as the phishing websites. Phishing is popular among attackers because it's easier to trick (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 Architecture Design Document?

The main objective of the Architecture design documentation is to provide the internal logic understanding of the flight fare prediction code. The Architecture design documentation is designed in such a way that the programmer can directly code after reading each module description in the documentation.

2. Architecture



3. Architecture Design

This project is designed to make an interface for the User to predict the rating of restaurant.

3.1. 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

3.2. 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.

3.3. Data Pre-processing

- 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.

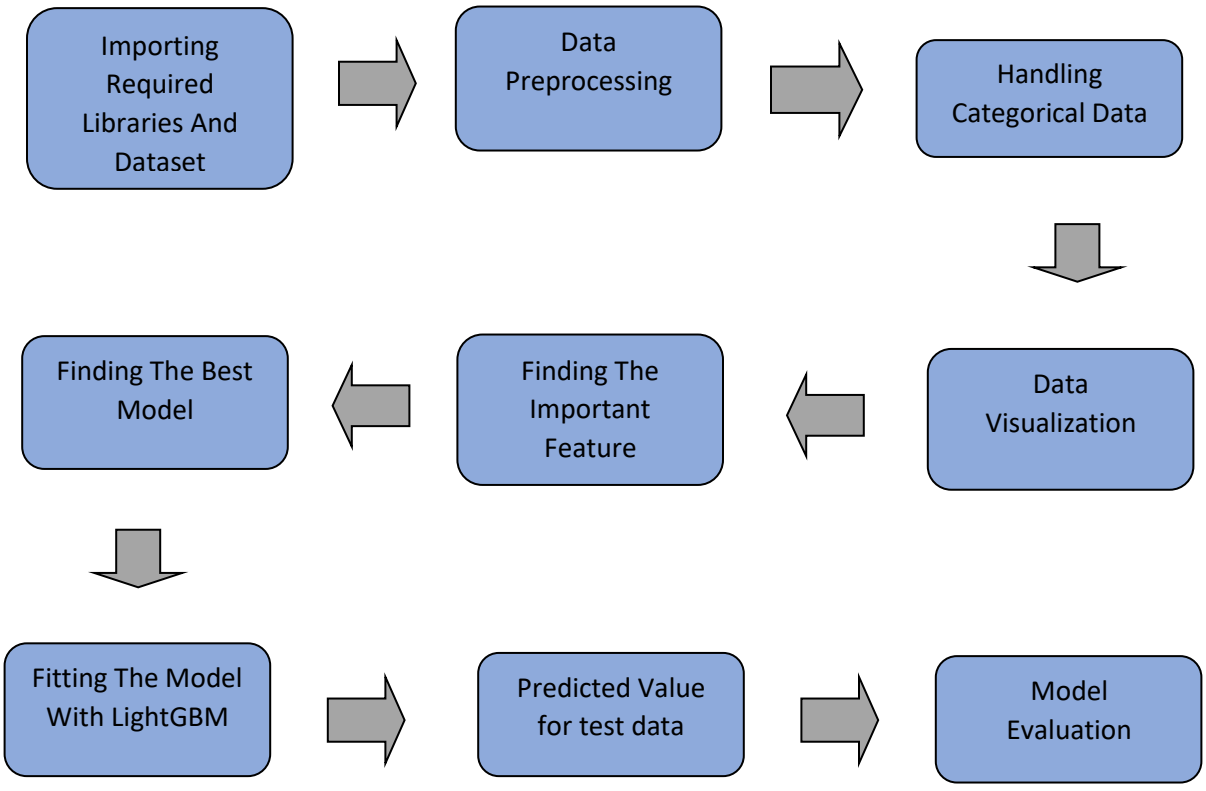
3.4. Feature Selection

The features are selected on ranking by LightGBM.

3.5. Model Creation

The Pre - processed info is now envisioned and drawn insights helps us to select the feature that improves the accuracy of the model. The info is randomly used for modelling with different machine learning algorithms to create a model to predict the Phishing website. After performing on

different algorithms, we use Random Forest Regression to create a model and then also perform Hyperparameter Tuning to improve the accuracy of the model.



3.6. 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.

3.7. Data From User

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

3.8. 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.

3.9. Data Validation

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

3.10. Rendering The Result

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

3.11. Deployment

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