

Phishing Attack Domain Detection

## Overview

This project focuses on solving the most common cyber attack called phishing. Phishing is a type of social engineering attack often used to steal user data, including login credentials and credit card numbers.

The purpose of this HLD document is to explain the high level details about the project. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

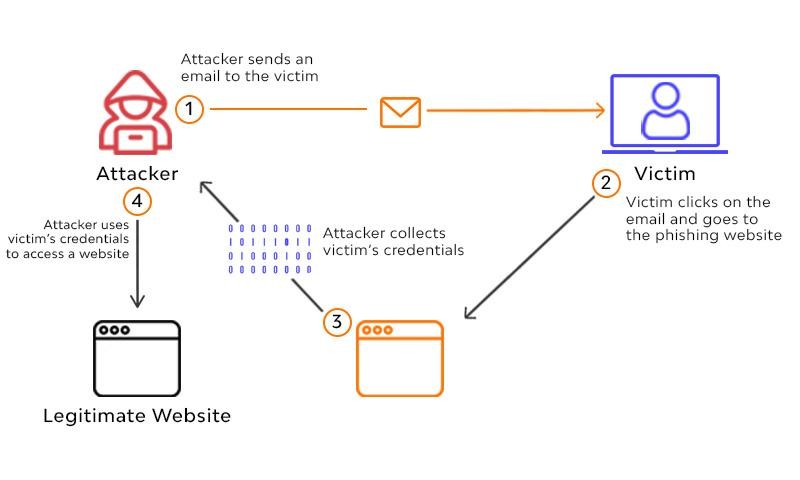
**Goals of HL***M***D**

1. **Explain the problem statement :** Explain the problem statement for the project and also explain how it affects the current situation.
2. **Technical Overview:** Give an of proposed solution and define all the requirements that will be used (including software & hardware) to build the system.



2

## Problem Statement

We live in the digital era. This has steadily changed the way you buy things, pay your bills, rent an apartment, watch a movie, and everything else. This project focuses on solving a most common cyber fraud called “Phishing”.

*Madc bQ Dccpcsm Mmatic*

Phishing is a type of social engineering attack often used to steal user data, including login credentials and credit card numbers.

It occurs when an attacker, masquerading as a trusted entity, dupes a victim into opening an email, instant message, or text message. The recipient is then tricked into **clicking a malicious link**, which can lead to the installation of malware, the freezing of the system as part of a ransomware attack or the revealing of sensitive information











## 1.Technical Overview

In this section we’ll discuss the technical solution for the above mentioned problem statement and also define some of the technologies we’ll be using to create the system.

## Proposed Solution

As we discussed above, most people either lack proper knowledge to identify a phishing attack or fail to take appropriate precautions, and any sort of cybersecurity training for company employees will only interrupt their daily workings.

Most phishing attacks happen when a user clicks a malicious URL, so our primary goal should be to **prevent the user from clicking any malicious**

**links**.

So in order to prevent such phishing attacks we propose a Machine Learning solution that can automatically analyze incoming malicious URLs and classify the links on its own rather than putting the Human as the last line of defense.

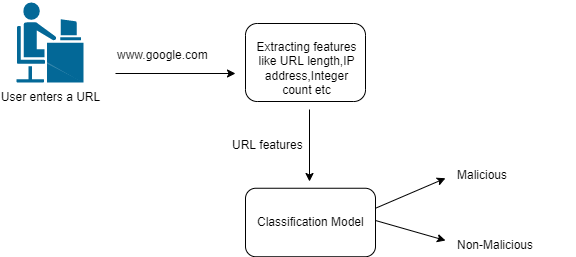
The end-goal machine learning system can be then used in web



10

## Classifier Architecture

The below Image shows a high level architecture of our trained classification model. Basically it takes a single URL from the user, Extracts the url’s features and then makes a prediction based on them whether the given URL is malicious or not.



*Madc bQ Dccpcsm Mmatic*

The prediction of the classifier model is high based upon the features of the given URL like its length,Integer count,IP address presence in domain etc.

This model will be designed to function as standalone so that it can be used in various applications like web browsers.search engines,social media sites etc.

**NOTE** : More details on the classification model in the LLD document.



11

## Data Requirements

In order to create a binary classification model which can optimally predict malicious phishing URLs,first we need to train the model on similar datasets.

This kind of data can be obtained from either public dataset or through scrapping suitable websites. Some of the properties that make a good dataset for our problem are the following :

* The dataset must contain an equal amount of positive and negative samples.
* The data must contain diverse features and characteristics with less sparsity among features.
* The data m*M*us*a*t c*d*on*c*tai*b*n b*Q*ot*D*h n*c*u*c*m*p*er*c*ica*s*l*m*and*M*ca*m*te*a*go*t*r*i*ic*c*al features in balance.
* The dataset must be verified, if possible contain real-life scenarios for optimal results and less artificially generated samples.

Some example datasets are listed below :

DatasetLink:<https://data.mendeley.com/datasets/72ptz43s9v/1>



12

## Technology Requirements

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, TensorFlow, Keras will be used to build the whole model.

*Madc bQ Dccpcsm Mmatic*

* PyCharm is used as an IDE.
* For visualization of the plots, Matplotlib,Seaborn and Plotly are used.
* Heroku is used for deployment of the model.
* Front end development is done using React & Css
* Python FastAPI is used for backend development.
* GitHub is used as a version control system.



13

## Conclusion

In this HLD document we skimmed over the problem statement of phishing attacks and also discussed our proposed solution. But this was just a high level summary for project managers, we’ll discuss the project in much more detail in the LLD document.

*Madc bQ Dccpcsm Mmatic*