SafeURL Inspector

Hemanth Kumar Reddy Tiyyagura
Masters in Computer Science
George Mason University
htiyyagu@gmu.edu
(703)-975-7183

Sampath Sai Yelleti
Masters in Data Analytics Engineering
George Mason University
syelleti@gmu.edu
(703)-477-7945

Chandra Kiran Viswanath Balusu
Masters in Computer Science
George Mason University
cbalusu@gmu.edu
(703)-309-4425

Anusha Bhavanam
Masters in Computer Science
George Mason University
abhavana@gmu.edu
(703)-501-4011

Siva Satyanarayana Raju Pusapati
Masters in Computer Science
George Mason University
sraju5@gmu.edu
(571)-318-1082

Poojasree Keerthipati
Masters in Data Analytics Engineering
George Mason University
pkeerthi@gmu.edu
(571)-567-6949

Oct 29, 2023

Team Name: Tech Titans

Abstract

The proliferation of counterfeit websites represents a growing concern in the digital landscape, where malicious actors aim to exploit the credibility and authority of institutional entities for illicit purposes. Such deceptive practices pose significant risks to citizens, businesses, and the government itself. This project addresses the crucial problem of online security and reliability by developing a comprehensive system for identifying fake URLs. It involves data preprocessing, feature engineering, and the utilization of machine learning algorithms, including Random Forest, Decision Tree, and K-Nearest Neighbors, to identify fraudulent URLs. The project seamlessly integrates a user-friendly front end for text-based URL input and a back end for extracting URL features as well as classifying the extracted URL into legitimate or fraudulent.To enhance the system's robustness, we added a second layer of authentication that uses extracted URL parameters from an SSL certificate .The front end provides real-time predictions from both methods and offers valuable insights into URL characteristics, enabling users to make informed decisions about the authenticity and potential risks associated with web resources. This project serves as a practical solution to enhance online security and user awareness by effectively identifying fake URLs.

1 Value Proposition

The specific problem that our project aims to address is the challenge of detecting fake urls in text messages,emails and the risks associated with the counterfit URLs.Counterfeit URLs are websites that replicate legitimate sites or attempt to deceive users for various malicious purposes, such as phishing, spreading malware, stealing personal information, or conducting other fraudulent activities.

The issue with this particular gap is that counterfeit URLs can cause significant harm to both individuals and organisations. Users who visit such websites are at risk of having their personal and financial information stolen, falling victim to scams, or having malware installed on their devices. Also, fake URLs can harm the reputation and credibility of the legitimate organisations and websites that they impersonate.

Stakeholders and Beneficiaries:

- **End Users:**
  Everyone who use the internet and might become targets of fake URLs will benefit from increased online security and awareness.
- **Government Authorities:**
  Cybersecurity and law enforcement agencies can use this solution to combat online fraud in order to safeguard citizens and businesses.
- **Businesses:**
  Online businesses can use this solution to protect their customers, reputation, and digital assets.
- **Cybersecurity Experts:**
  Professionals in the field of online security who can benefit from the system's insights and contribute to its improvement.
- **Website Owners:**
  who own legitimate websites and can use this solution to monitor and report potential phishing or fraudulent websites imitating their brand.
- **Internet Service Providers (ISPs):**
  Organizations that can utilize this solution to enhance the safety of their networks and protect their users from malicious websites.
- **Educational Institutions:**
  Educational Institutions can incorporate this system into their cybersecurity curriculum to educate students about identifying and avoiding fake URLs.

2 Methodology

This project describes the multi-staged implementation of fake url identification utilizing the Machine Learning models for classifying URLs into fake or legit,Selenium for converting short length URL to long length URL and SSL to extract certificate parameters.The first stage entails converting short length URL to long length URL as it is difficult to extract features from short length URL.With the help of Selenium we can get the redirected URL and verify the SSL certificate of the redirected URL whether it is valid or not using the parameters like Data of Expiry of the website and domain name. Redirected URL with parameters extracted is then passsed to the Machine Learning Model for classification task.The outputs from both the implementations are then used to verify the URL legitimacy.Figure 1 shows the stages in the proposed approach.

2.1 Dataset

Dataset is taken from Kaggle website.The dataset includes 11430 URLs with 87 extracted features.Features are from structure of URLs and from content of their pages.The dataset is balanced,with 50% phishing and 50% legitimate URLs.

2.2 Selenium

Most of the text messages and emails contains short length URL and we cannot extract features from the short length URL.For this purpose we need to convert it into long length URL for analysis purpose.Selenium takes the extracted URL from text or email and gives the redirected URL(long length URL)

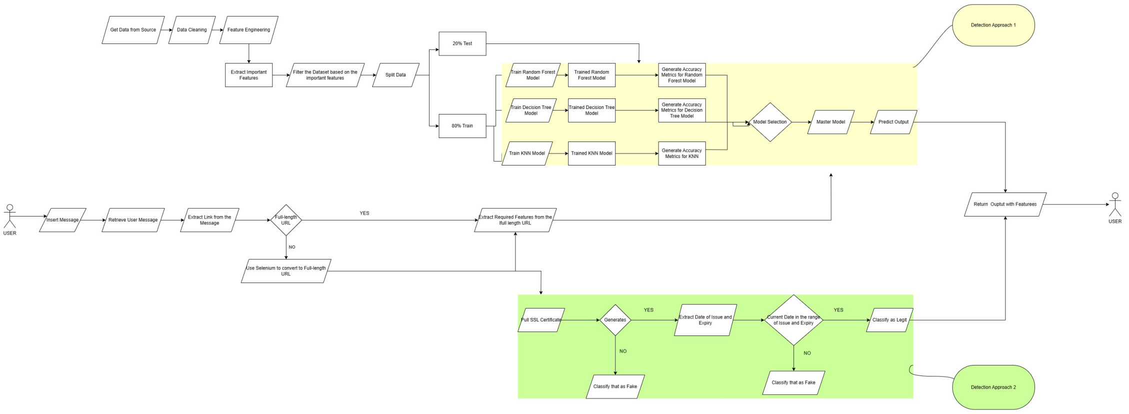
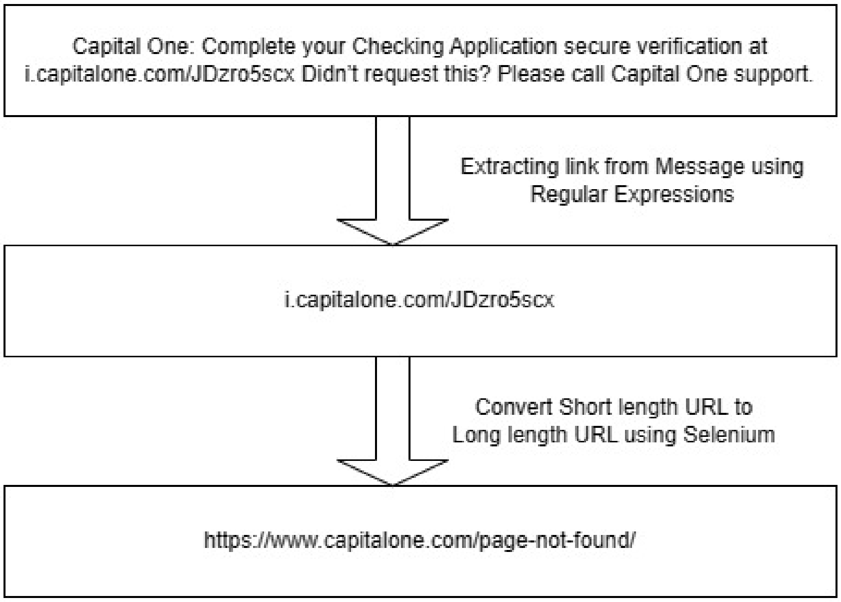

Get Data from Source
Data Cleaning
Feature Engineering
Extract Important Features
Filter the Dataset based on the important features
Split Data
20% Test
80% Train
Train Random Forest Model
Trained Random Forest Model
Generate Accuracy Metrics for Random Forest Model
Train Decision Tree Model
Trained Decision Tree Model
Generate Accuracy Metrics for Decision Tree Model
Train KNN Model
Trained KNN Model
Generate Accuracy Metrics for KNN
Model Selection
Master Model
Predict Output
Detection Approach 1
USER
Insert Message
Retrieve User Message
Extract Link from the Message
Full-length URL
YES
NO
Extract Required Features from the lfull length URL
Use Selenium to convert to Full-length URL
Return Ouptut with Featurees
USER
Pull SSL Certificate
Generates
YES
NO
Extract Date of Issue and Expiry
Current Date in the range of Issue and Expiry
YES
NO
Classify as Legit
Classify that as Fake
Classify that as Fake
Detection Approach 2

Figure 1: Proposed Approach


Capital One: Complete your Checking Application secure verification at i.capitalone.com/JDzro5scx Didn't request this? Please call Capital One support.

Extracting link from Message using Regular Expressions

i.capitalone.com/JDzro5scx

Convert Short length URL to Long length URL using Selenium

https://www.capitalone.com/page-not-found/

Figure 2: Regex and Selenium workflow

2.3 SSL Certficate

Using OpenSSL command we retrive SSL certificate of the URL.It has the following parameters:

- Issuer
- Expiry Date
- Domain Name
- Alternate Names

2.4 Machine Learning Models for Classification

In the early stage Feature Engineering is performed on the dataset by using Correlation Matrix, the output from the correlation matrix helped us to filter out the most correlated features with the output. This helped us bring down the features from 87 to 18 which can be considered as most important features. We later split the data into two proportions of size 80% and 20% into Training and Testing using ScikitLearn. Machine Learning models like Decision Tree, Random

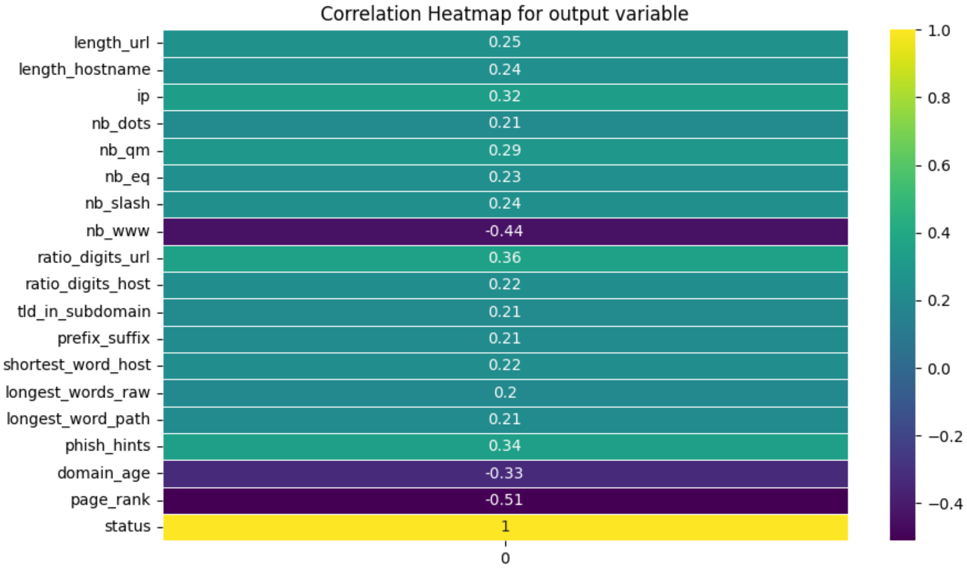
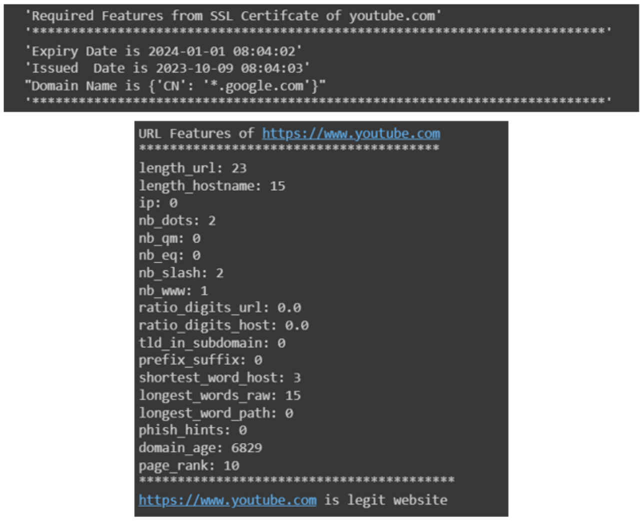


Figure 3: Correlaation Matrix

Forest and KNN Classifier are build and trained using the train data. Test data is then used to find out the metrics that define the model performance.The metrics are then analyzed and best model is selected from the three models. Random Forest turned out to be the best model for the data. For the second way of classification we retrieve the SSL certificate of the URL and retrieve the required features like issue date of certificate and expiry date of certificate, we compare the current date with the expiry date and issue date, if the current date lies in between issue date and expiry date, then we can consider the website as legit. From the front end point of view the message is given by the user and the message is parsed such a way that URL link from the message is retrieved. Later we use Selenium to convert the short format URL that is extracted into long format URL. This long format URL is examined to extract the required features into Method 1(ML Model) of Classification and Method 2(SSL Certificate) of Classification.



```
'Required Features from SSL Certifcate of youtube.com'
'****************************************************************************'
'Expiry Date is 2024-01-01 08:04:02'
'Issued  Date is 2023-10-09 08:04:03'
"Domain Name is {'CN': '*.google.com'}"
'****************************************************************************'
```

```
URL Features of https://www.youtube.com
***************************************
length_url: 23
length_hostname: 15
ip: 0
nb_dots: 2
nb_qm: 0
nb_eq: 0
nb_slash: 2
nb_www: 1
ratio_digits_url: 0.0
ratio_digits_host: 0.0
tld_in_subdomain: 0
prefix_suffix: 0
shortest_word_host: 3
longest_words_raw: 15
longest_word_path: 0
phish_hints: 0
domain_age: 6829
page_rank: 10
*****************************************
https://www.youtube.com is legit website
```

Figure 4: SSL Certificate and Random Forest Outputs

3 Potential Applications

Benefits for the End Users:

- **Boost Online Safety:**
  Users can browse the web with greater confidence, knowing that they are less likely to come across bogus websites or become victims of scams.
- **Reduce Risks:**
  The solution reduces the danger of identity theft and financial loss by lowering the chance of personal and financial information theft.
- **Save Time and Effort:**
  Users will save time and stress by not having to wade through potentially fraudulent websites because the system delivers real-time assessments.
- **Increase Awareness:**
  It provides users with insights into URL properties, allowing them to be more discerning and informed as they navigate across the web.
- **Strengthen Trust:**
  Users can have confidence that they are engaging with trustworthy websites, which promotes trust in e-commerce, online services, and information sources.

Strengths of the Solution:

- **Comprehensive Approach:**
  Our solution combines data preprocessing, feature engineering, and a variety of machine-learning techniques to provide a robust way of detecting fraudulent URLs.
- **SSL Certificate Authentication:**
  Reliability and fraud detection are improved by adding an additional layer of authentication with SSL certificate specifications.
- **User Insights:**
  By gaining insightful knowledge about URL attributes, users are better equipped to choose web resources.
- **Multi-Stakeholder Benefits:**
  The system offers a comprehensive approach to online security and serves a wide range of stakeholders, including regulatory agencies and end users.
- **URL Extraction from Text or Email:**
  Existing websites does not extract URL from text where only URL should be given in the text area.But our solution can take text or email data and URL is extracted from it which makes more innovative from other websites.

Potential Weaknesses,Vulnerabilities and Constraints

- **Data Quality:**
  The training data's quality has a major impact on how accurate the solution is. False positives or false negatives may result from biased or inaccurate data.
- **Evolving Threats:**
  Hackers adapt new ways to generate false URLs arise on a regular basis, which may not be covered by existing models.
- **Resource Intensive:**
  Machine learning models and real-time processing are resource-intensive, which may result in longer response times or scalability concerns.
- **SSL Certificate Availability:**
  Some of the fake websites have SSL certificates, limiting the effectiveness of the second layer of authentication.

Mitigation

- **Data Quality Assurance:**
  Implement rigorous data quality checks and continuous monitoring to minimize the impact of inaccurate data. Regularly update the training dataset to reflect evolving threats.
- **Adaptive Models:**
  Update and refine machine learning models on a regular basis to respond to new threats. Use anomaly detection tools to spot new attack trends.
- **Resource Optimization:**
  Investigate cloud-based technologies and parallel processing to optimize resource utilization while maintaining real-time predictions and scalability.
- **Fallback Mechanisms:**
  When SSL certificates are unavailable, develop alternate techniques for URL authentication, such as domain reputation analysis.

Feasibility

- **Data and Model Building:**
  The utilization of a Kaggle dataset comprising 84 features, followed by data cleaning and feature engineering to identify the top 20 important features, provides a good foundation for model building. This method provideds a thorough understanding of the dataset and enhanced the feature set for more accurate predictions.
- **Model Selection and Evaluation:**
  Training different classifiers like Random Forest, Decision Tree, and KNN—made it possible to undertake a comparative study to find the best-performing model. The accuracy measures used to evaluate these models on test data provided a clear assessment of their predictive ability. The methods used in model evaluation provides assurance in the robustness of the chosen model.
- **Back-End Implementation:**
  The core functionality of the back-end system, which includes extracting URLs from user input messages, extracting necessary features, and using these features in machine learning models, demonstrates a well-organized implementation strategy. Additionally, the use of SSL certificate parameters as an additional prediction method demonstrates an accurate approach to validation.
- **Front-End Implementation**
  A user-friendly experience is ensured by the UI, which includes a text box for user input and the presentation of the results of the back-end processing. Displaying the predictions and URL characteristics in a logical order improves the user's awareness of the analysis.

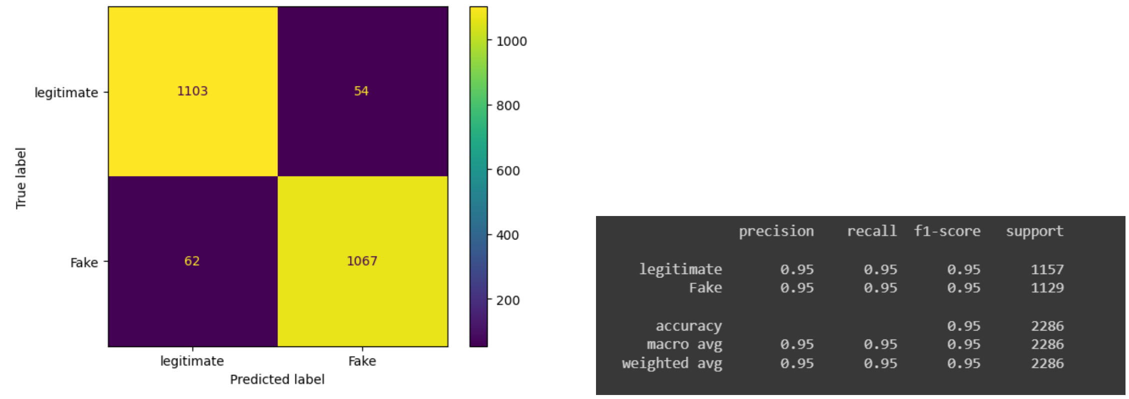
Impact and Metrics

- **Impact:**
  Impact of our solution is to significantly reduce financial loss,prevent information theft to enhance user data security and privacy,user trust and confidence on the legitimate websites.
- **Metrics:**
  We used Accuracy,Precision,Recall as metrics to measure the impact of our solution.Accuracy of our model on test data is 95%

4 Assumptions and Limitations

Assumptions

- **Data Quality:**
  The initial Kaggle dataset is assumed to be of sufficient quality and to represent a diverse set of URL examples. The model's success is dependent on the quality and representativeness of the data.
- **Feature Importance:**
  The assumption was that the top 20 important features selected through feature engineering are the most relevant for URL detection. The accuracy of this feature selection process determines the model's effectiveness.



```
              precision    recall  f1-score   support

  legitimate       0.95      0.95      0.95      1157
        Fake       0.95      0.95      0.95      1129

    accuracy                           0.95      2286
   macro avg       0.95      0.95      0.95      2286
weighted avg       0.95      0.95      0.95      2286
```

(a) Confusion Matrix

(b) Classification Report

Figure 5: Metrics

- **Feature Extraction Accuracy:**
  The extraction of features from URLs and SSL certificates was assumed to be precise and reliable. Variations in URL structures and SSL configurations, on the other hand, may cause inconsistencies in feature extraction, potentially affecting the model's predictive accuracy.

Limitations

- **Real-Time Updates:**
  The solution does not take into account model updates in real time. The characteristics and patterns of fake URLs may change over time, and the model should be retrained on a regular basis to adapt to evolving threats.
- **Scalability:**
  In real-world scenarios, the system may be required to process a large number of incoming URLs, and the infrastructure's scalability may be a bottleneck.
- **Feature Extraction Accuracy:**
  The extraction of features from URLs and SSL certificates was assumed to be precise and reliable. Variations in URL structures and SSL configurations, on the other hand, may cause inconsistencies in feature extraction, potentially affecting the model's predictive accuracy.
- **Fake SSL Certificate:**
  It is critical to understand that some malicious actors create fake websites with valid SSL certificates. This poses a significant limitation to the solution. Such certificates can be obtained in a variety of ways, including through free or low-cost certificate authorities, making it more difficult to rely solely on SSL certificate information for URL legitimacy assessment. As a result, the solution may still have difficulty in accurately identifying fake websites that use SSL certificates.

5 Operational Requirements

Programs Used:

- **Frontend:**
  Angular,Particle JS, Bootstrap
- **Backend Restful Service:**
  FastAPI With Uvicorn ASGI Server,Databases(module) for Database Async Connections,SQLAlchemy - ORM,Selenium - Handling Redirects,Openssl - Getting and Handling SSL Certificates,Pydantic - Restapi response models,pyTelegramBotAPI - For sending Telegram messages,Scikit Learn - Random Forest Classifier,python whois - Get the website age,OpenPage Rank API - For getting page rank
- **Android Application:**
  Kotlin,Android Studio

Data Required:

- **For Android Application Users:**
  User will need to provide a Telegram Bot API Key and his telegram user ID to get the detection Messages
  Android Application should be given Necessary Permission to Read Messages and The access to internet
  Android Application dont have an interface, its a backend service
  Open Page Rank API Key is Required to find Page Rank

6 Implementation Requirements

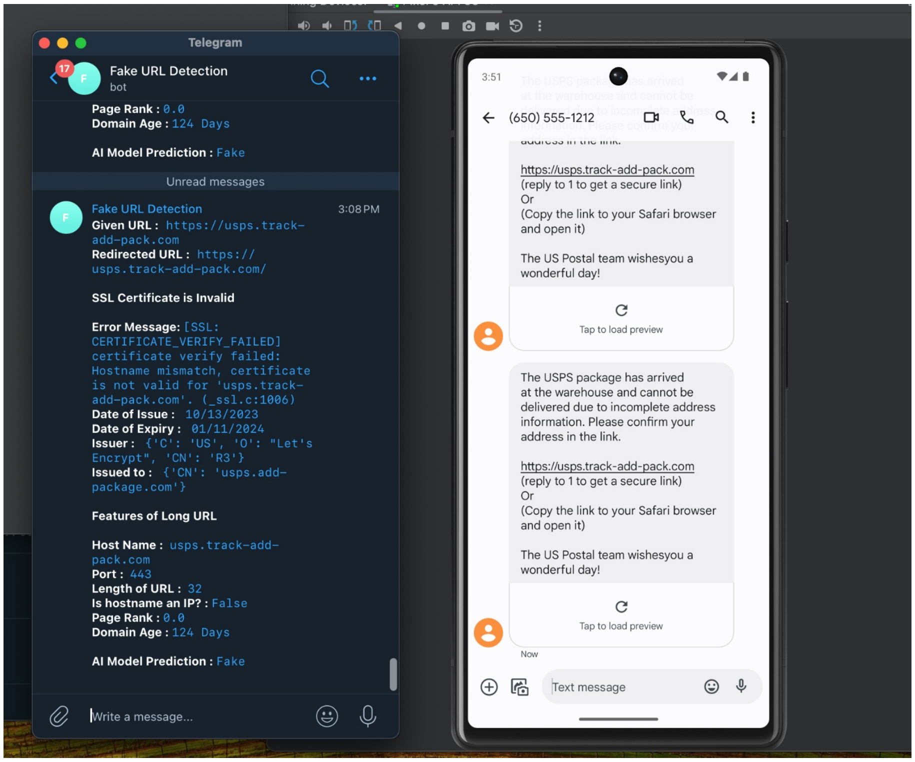
Policies: End User needs to accept permissions to read messages by the application when installing application in mobile phone
Resources: End User needs to have telegram,mobile phone to access app and website
Budget: Free of cost for end user to use app and website

7 Next Steps

In order to improve model we need more training data.More Data can be taken from Government Institutions,CyberSecurity Team and Analysts.Collected data can be given to the model to learn more and more about fake URLS to be more precise.We also plan to deploy our website in cloud to make it open for all the public.We will create feedback form where users can report false negatives and false positives which can be used to fine tune the model.We will develop continuous learning for the model to adapt to new emerging threats.We will be adding glossary where users can understand why the URL is fake or legit from the predictions and the features extracted from the URL.Deploying our app on online platforms and social media.

8 Estimated Budget

API Calls for Page Rank Calculation- 500 USD(per month)
Cloud Hosting - 20 USD(per month)
Hosting App in Play Store - 25 USD(one time)
Hosting App in App Store - 99 USD(per year)



Page Rank : 0.0
Domain Age : 124 Days

AI Model Prediction : Fake

Fake URL Detection 3:08 PM
Given URL : https://usps.track-add-pack.com
Redirected URL : https://usps.track-add-pack.com/

SSL Certificate is Invalid

Error Message: [SSL: CERTIFICATE_VERIFY_FAILED] certificate verify failed: Hostname mismatch, certificate is not valid for 'usps.track-add-pack.com'. (_ssl.c:1006)
Date of Issue : 10/13/2023
Date of Expiry : 01/11/2024
Issuer : {'C': 'US', 'O': "Let's Encrypt", 'CN': 'R3'}
Issued to : {'CN': 'usps.add-package.com'}

Features of Long URL

Host Name : usps.track-add-pack.com
Port : 443
Length of URL : 32
Is hostname an IP? : False
Page Rank : 0.0
Domain Age : 124 Days

AI Model Prediction : Fake

address in the link.

https://usps.track-add-pack.com
(reply to 1 to get a secure link)
Or
(Copy the link to your Safari browser and open it)

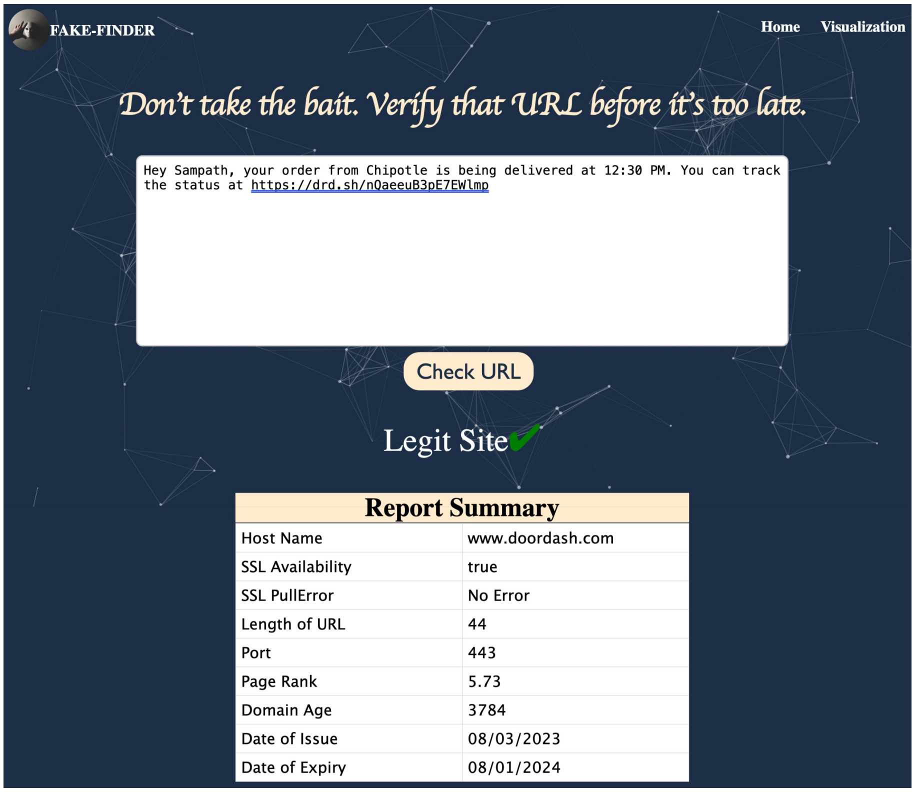
The US Postal team wishesyou a wonderful day!

The USPS package has arrived at the warehouse and cannot be delivered due to incomplete address information. Please confirm your address in the link.

https://usps.track-add-pack.com
(reply to 1 to get a secure link)
Or
(Copy the link to your Safari browser and open it)

The US Postal team wishesyou a wonderful day!

Figure 6: Android App Result



Hey Sampath, your order from Chipotle is being delivered at 12:30 PM. You can track the status at https://drd.sh/nQaeeuB3pE7EWlmp

Report Summary	
Host Name	www.doordash.com
SSL Availability	true
SSL PullError	No Error
Length of URL	44
Port	443
Page Rank	5.73
Domain Age	3784
Date of Issue	08/03/2023
Date of Expiry	08/01/2024

Figure 7: Website Result