Phishing Domain Detection (PS Code - SIH 1454)
By Team Chakde India

Methodology-

The URL of the unknown website is checked at the beginning of the online phase to determine whether it has been detected before. If it has been detected, it is skipped directly. Otherwise, it is entered into the visual similarity to compare its domain and contours. The operations are shown in **Fig 1**.

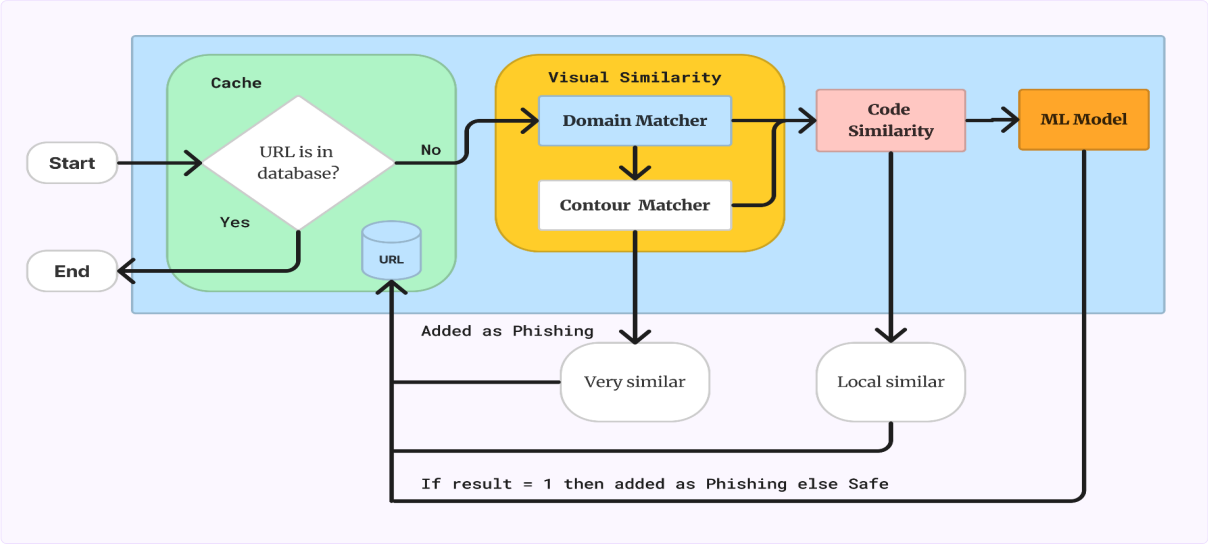


Fig 1. Operations in Online Phase

The domain of the entered url is compared with the database of url by extracting keywords and the top three urls which have least Levenshtein distance and longest common subsequence are returned as shown in **Fig 2**.

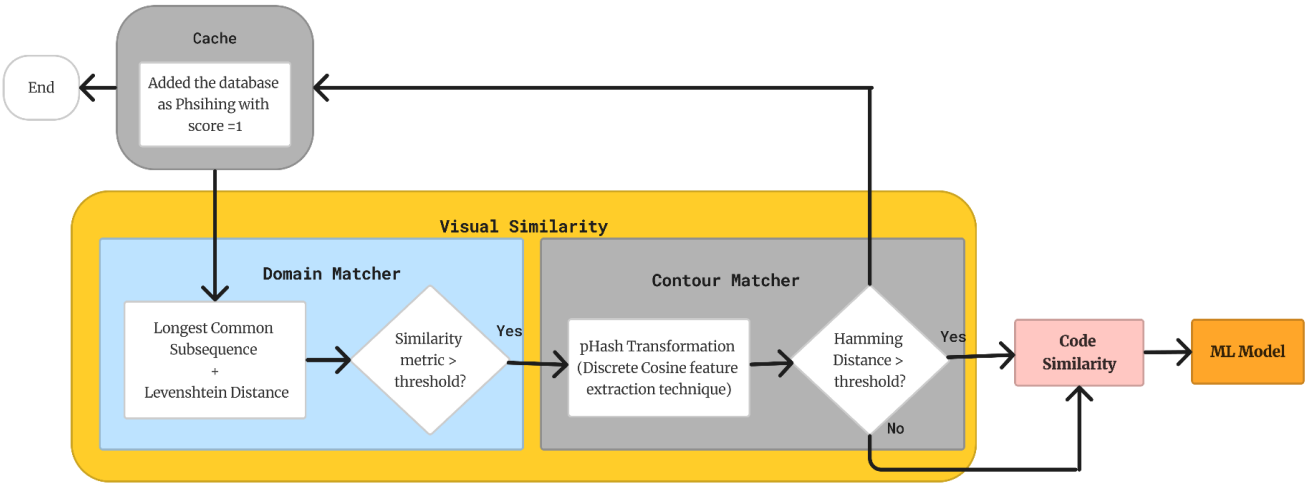


Fig 2. Visual Similarity

Next, the screenshots of these urls is captured and converted into hash using pHash and the hamming distance is calculated between the hash value of the target url and the matched domain and if the hamming distance value is less than 15 it is phishing domain and stored in database with score = 1 otherwise it is passed through code similarity.

Now if the visual similarity is not similar, the matched domains are passed through code similarity (**Fig 3**) where if hamming distance is more than 90% that url is marked as phishing and stored in database else it is passed through model

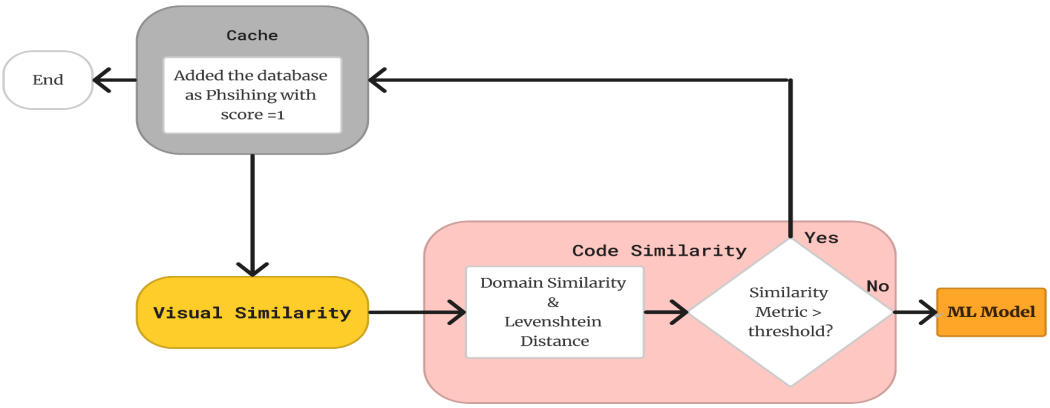


Fig 3. Code Similarity

If the code similarity tells that the html and dom structure of two urls are different then the url is lastly evaluated on the basis of model. The ml model is the ensemble of 4 classifiers (**Fig 4**) and gives the testing accuracy of 93%. The final result is then displayed on the website of SecurePhish.

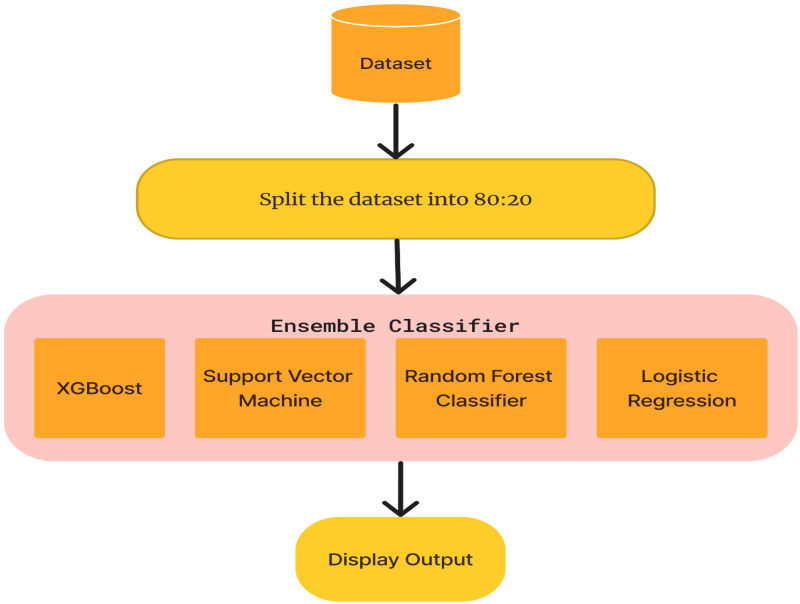
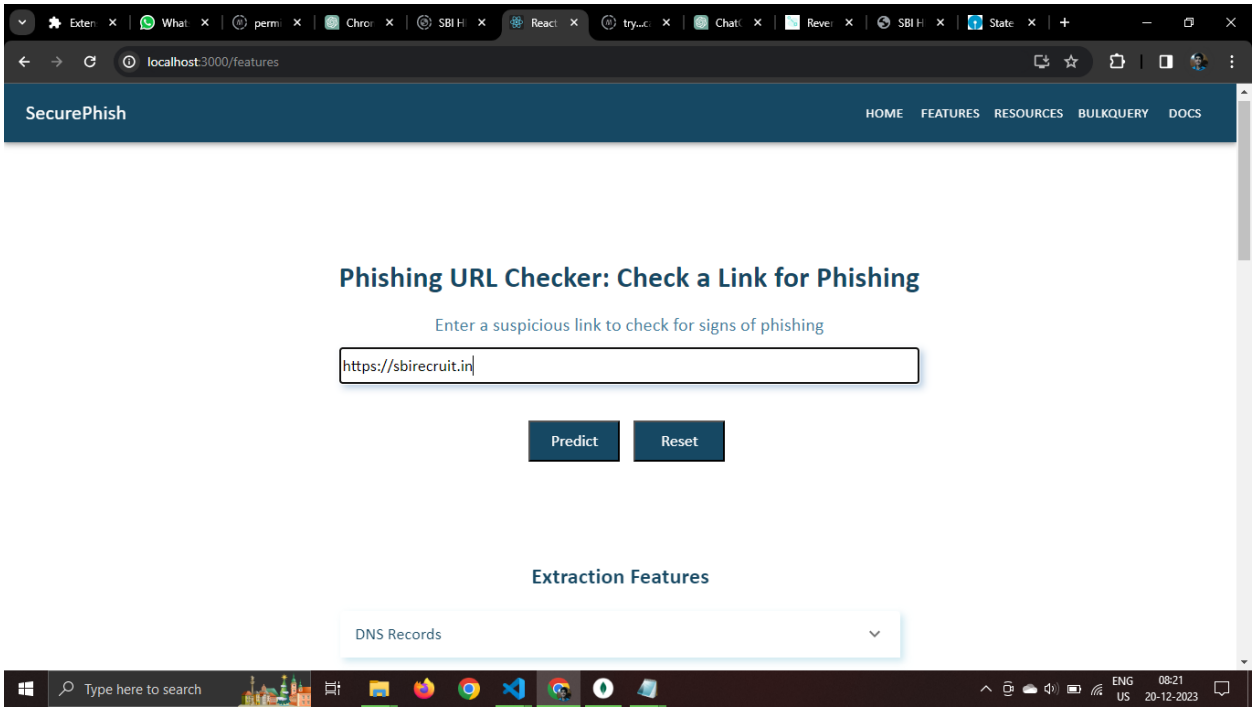
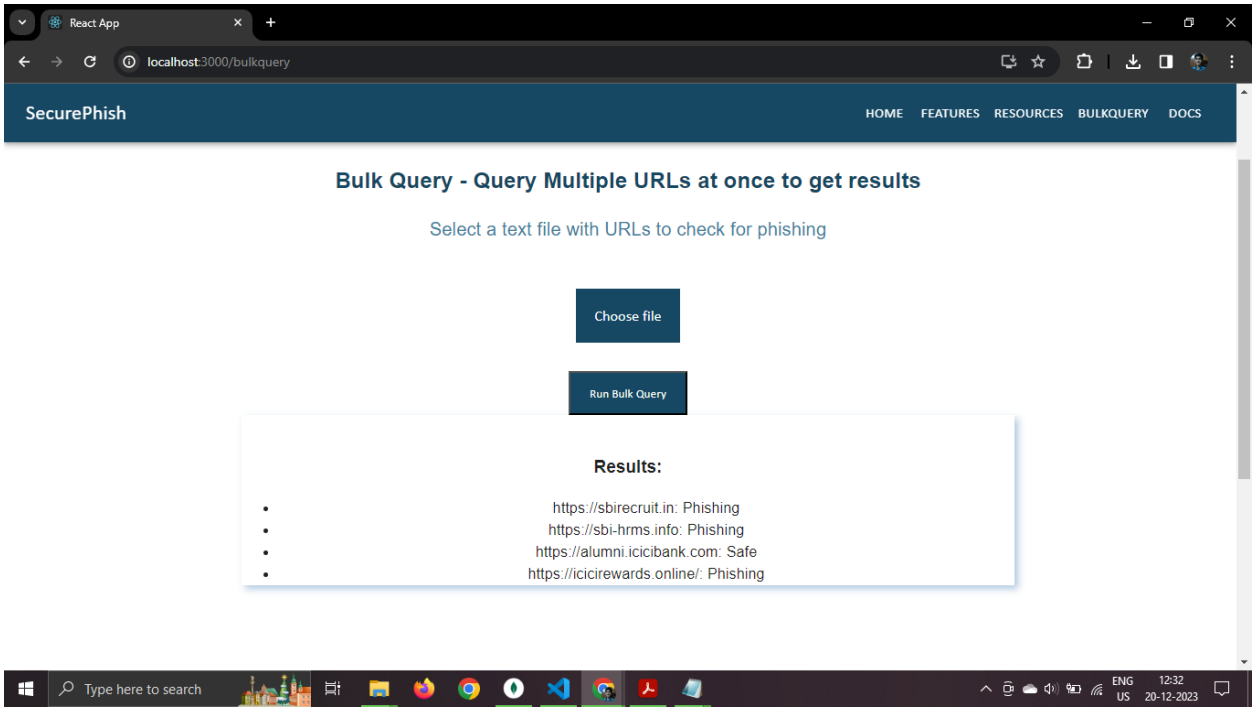


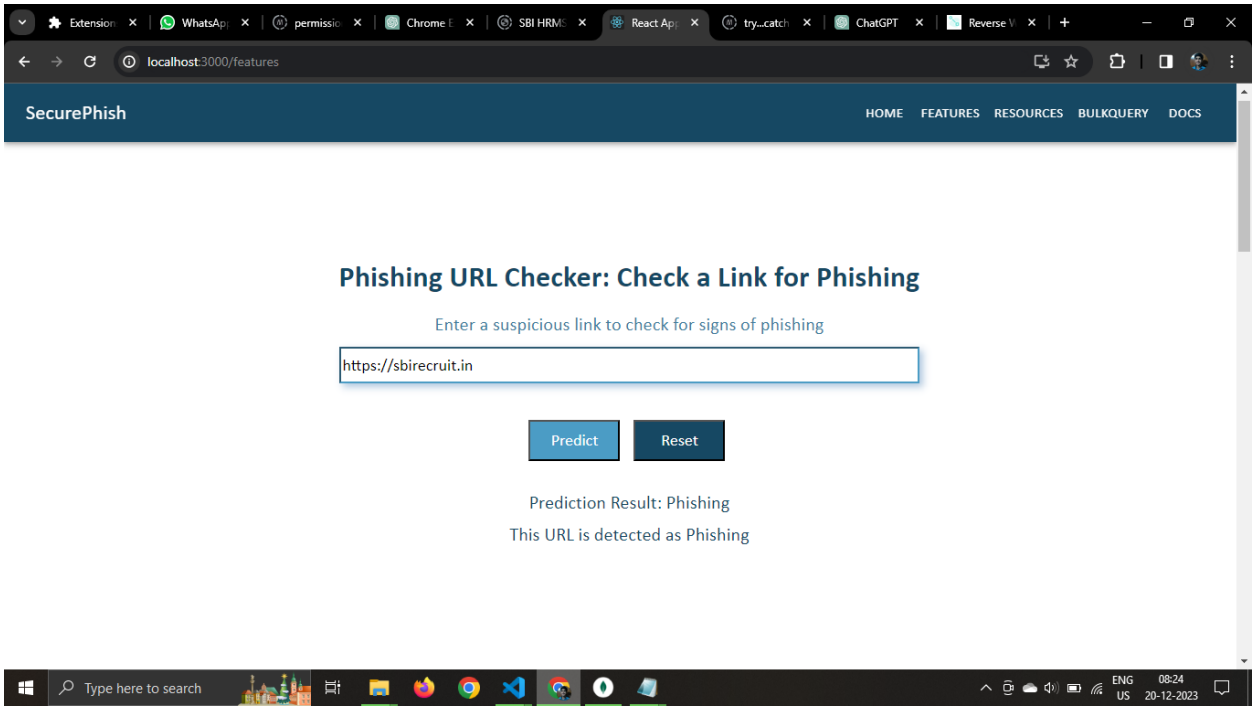
Fig 4. Model Implementation

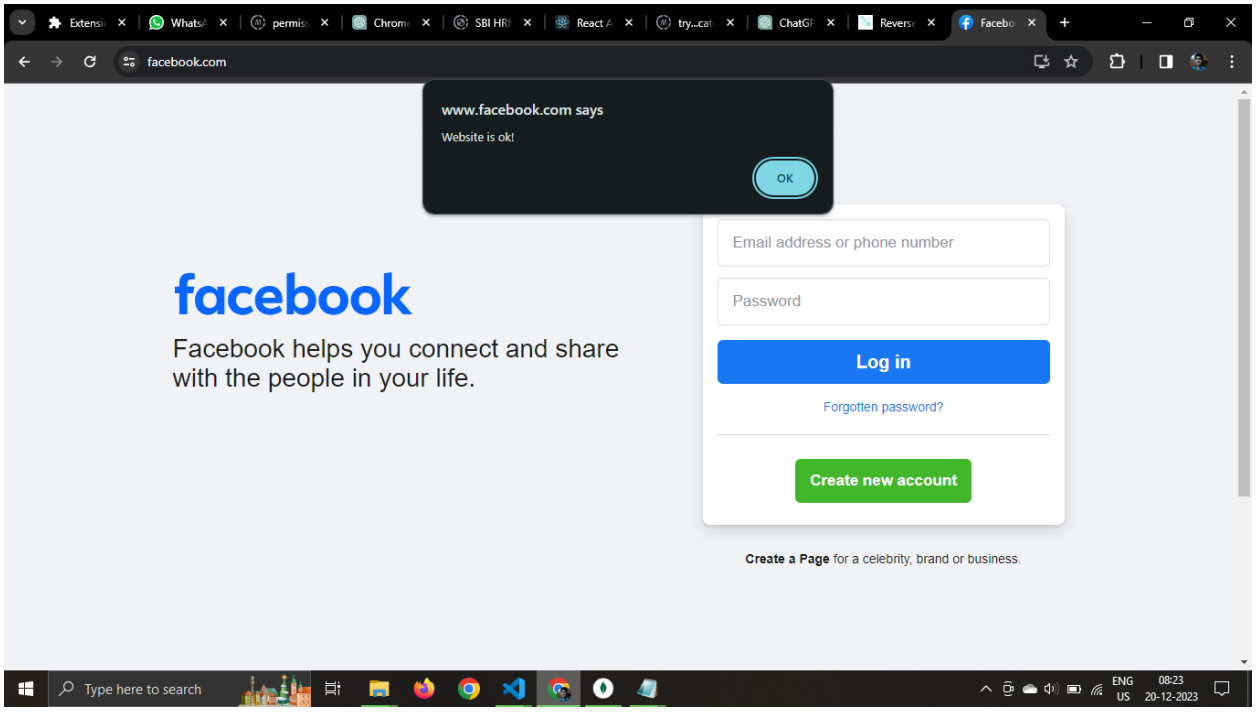
Key Features of SecurePhish-

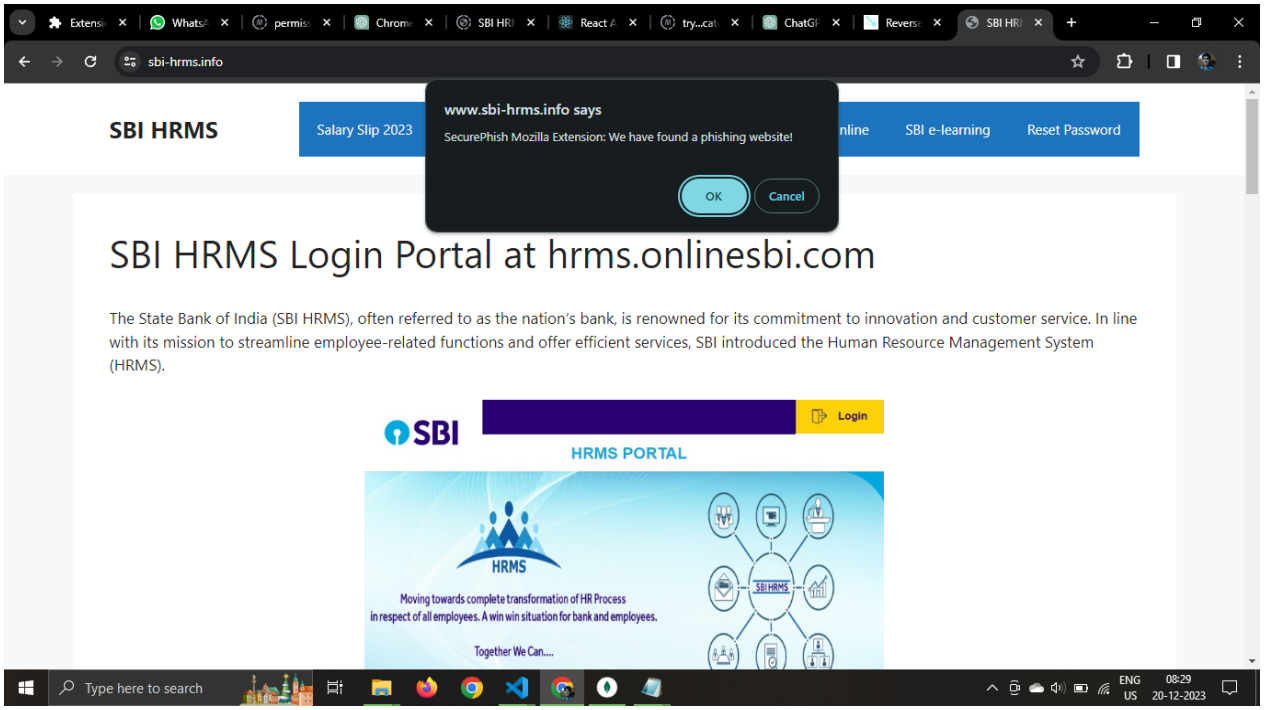
- Extraction of all subdomains of a domain using SubFinder, Findomain, Amass and alive subdomain list using HTTPx
- ML model trained on 18 features which include lexical, code and domain based features
- UserFriendly Frontend Website for our Product SecurePhish






localhost:3000/features
SecurePhish
HOME
FEATURES
RESOURCES
BULKQUERY
DOCS
Phishing URL Checker: Check a Link for Phishing
Enter a suspicious link to check for signs of phishing
https://sbirecruit.in
Predict
Reset
Prediction Result: Phishing
This URL is detected as Phishing
Type here to search
ENG
US
08:24
20-12-2023


facebook.com
www.facebook.com says
Website is ok!
OK
facebook
Facebook helps you connect and share with the people in your life.
Email address or phone number
Password
Log in
Forgotten password?
Create new account
Create a Page for a celebrity, brand or business.
Type here to search
ENG
US
08:23
20-12-2023

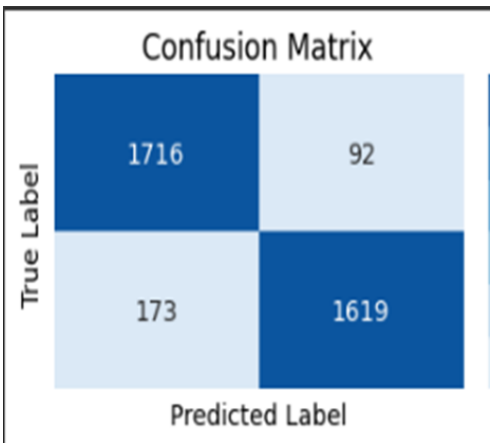


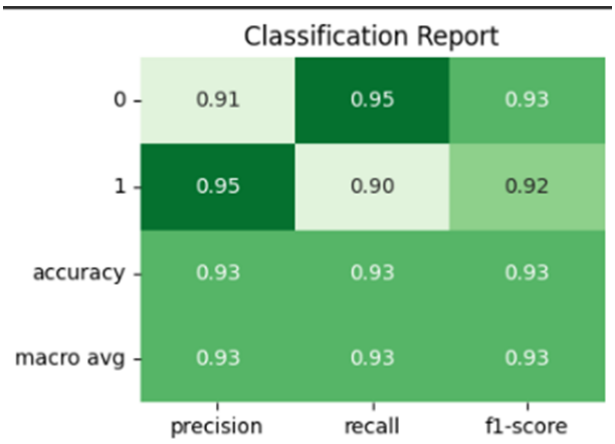
- Chrome and Firefox browser extensions
- Detection of Individually Queried URLs
- Bulk Query of a list of URLs - Input as a text file, and results are displayed on website and downloaded automatically as a Results text file
- Optimization Mechanism at the backend to ensure speed and efficiency

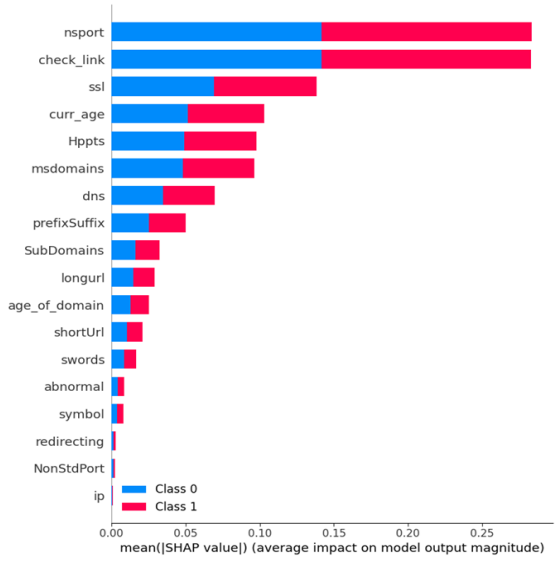
Metrics of our Algorithm-

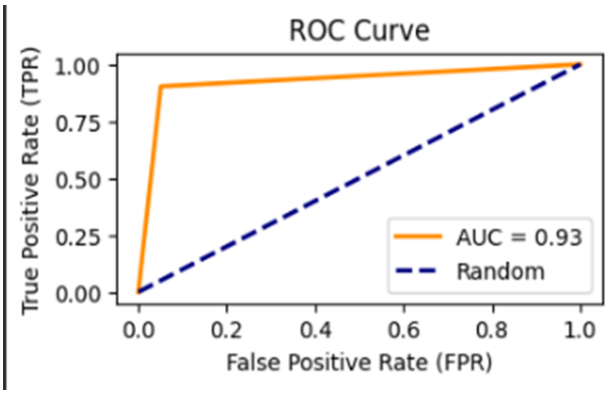
- Training Accuracy: 0.950
- Test Accuracy: 0.93
- ROC Curve Area: 0.93
- False positive rate: 0.025
- False negative rate: 0.0486
- F1 Score: 0.93
- Recall: 0.93

Screenshots of Metrics of ML Model-









Future enhancement -

- Keyword based domain priority
- Fetch and analyze URLs that were registered in the past or recently using ReverseWhoIs Lookup
- DNSDB API can be used to find historical registered domains
- Creating an interface for organizations using our product to alert them/display if any phishing urls of the original domains are found
- Scaling the project to handle large amount of requests by deploying on cloud and using multiple servers and load balancers