Large Application Practicum (CS308)
Project - WhtThePhish!!
Final Report
Group - 4

[Github Repository]

Members:

Rohan Raj Kansal (B19108)
Prashant Kumar (B19101)
Priyanshu Shubham (B19189)
Sourav Sehgal (B19059)
Shashwat Singh (B19056)
Ravi Kumar (B19191)
Pritish Chugh (B19187)

Acknowledgement

We express our deep gratitude and sincere thanks to the Respected Dr. Varun Dutt, our instructor of Large Application Practicum course for his encouragement and provided facilities for this project. We sincerely appreciate his generosity by taking us into his fold for which we shall remain indebted to him.

We extend our appreciation to TA Megha Sharma, our mentor for the project who guided us to the successful completion of this project. We take this opportunity to express our deep sense of gratitude to her invaluable guidance, ongoing encouragement, enormous motivation, which has sustained our efforts at all the stages of project development.

Table Of Contents:

Abstract

Phishing is a very prevalent issue in today's cyber world. It can have both financial and personal consequences. With the increase in Internet users every day, attacks continue to become more and more sophisticated and the advanced ones can only be detected if people carefully check URLs – be it in messages or in the address bar of the web browser. We have developed a game-based web app – "WhtThePhish" – to make people aware about various phishing techniques incorporated by attackers while creating phishing URLs; i.e. educating them to distinguish between trustworthy and untrustworthy websites by analysing their URLs. As the user progresses through the game new information is provided to him about various types of spoofing URLs on every level and then the understanding is tested in a playful manner. Various teaching and learning strategies are incorporated while making the game so that the user can understand the tricks involved and can demystify them if he/she ever encounters such a URL.

Introduction

Phishing is a fraudulent practice of creating false websites purporting to be from reputable companies in order to induce individuals to reveal personal information, such as passwords and credit card numbers. The financial benefit of Phishing is an incentive for phishers to keep luring victims into disclosing their sensitive information.

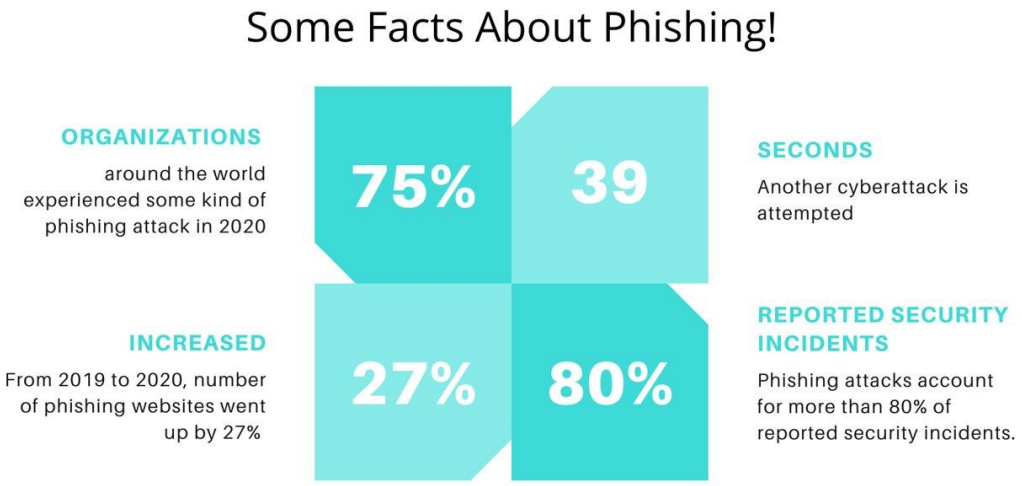

Fig 1. Some Facts About Phishing

Our goal is to develop a new webapp based game – "WhtThePhish" – an anti-phishing education web app that addresses these issues to provide more sophisticated knowledge on how to properly check URLs – be it in messages or in the address bar of the web browser. The detection of phishing on mobile browsers is complicated because in many cases address bars disappear and even if shown only parts of the URL are visible.

Technology Stack

1. **Frontend**: To make the user interface for our WebApp we have used the mentioned technologies,
    a. **HTML -** To make the structure of the website we have used HTML.
    b. **CSS -** To apply styling to various elements of the webpage we have used CSS
    c. **Javascript -** To apply transition and responsiveness to the website we have used Javascript and one of it's frameworks - **JQuery.**
2. **Backend**: To make the backend server for our WebApp we have used the mentioned technologies,
    a. **Flask -** We Have used the Python framework, Flask to create our server, with various libraries to get form information, PyMongo to connect to the Database etc.
3. **Language**: Python language was used in the backend.
4. **Database**: To store user information and question information we have used **MongoDb** and have created the mentioned Models,
    a. **User -** This collection contains the information about the user and it includes,
        i. **name :** name of the user
        ii. **email :** email of the user
        iii. **password :** hashed password of the user
        iv. **stars :** total stars the user has achieved so far
        v. **_id :** the unique id of the document
        vi. **level1 :** stars earned in level1
        vii. **level2 :** stars earned in level2
        viii. **level3 :** stars earned in level3
        ix. **level4 :** stars earned in level4
        x. **level5 :** stars earned in level5
        xi. **level6 :** stars earned in level6
        xii. **level7 :** stars earned in level7
    b. **Question -** This collection contains the information about the questions and it includes,
        i. **url :** URL of the website
        ii. **url_brand :** The company to which the URL belongs
        iii. **phishing :** is this URL phishing (1) or not (0)
        iv. **_id :** unique id for each document in this collection
        v. **category :**
5. **Version Control**: We have used the GIT as our version control system, so that every member of our group can contribute to the project without facing much difficulties.

Plan for developing the application

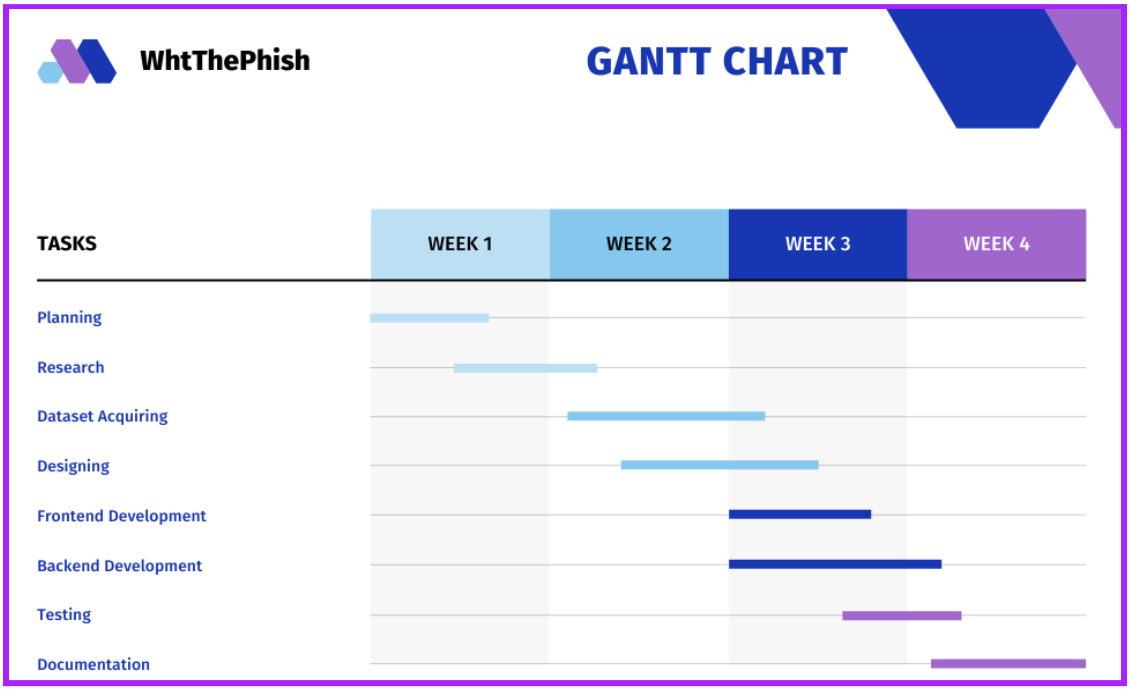

Fig. 2 Gantt Chart Showing Plan to complete the work.

Database Models

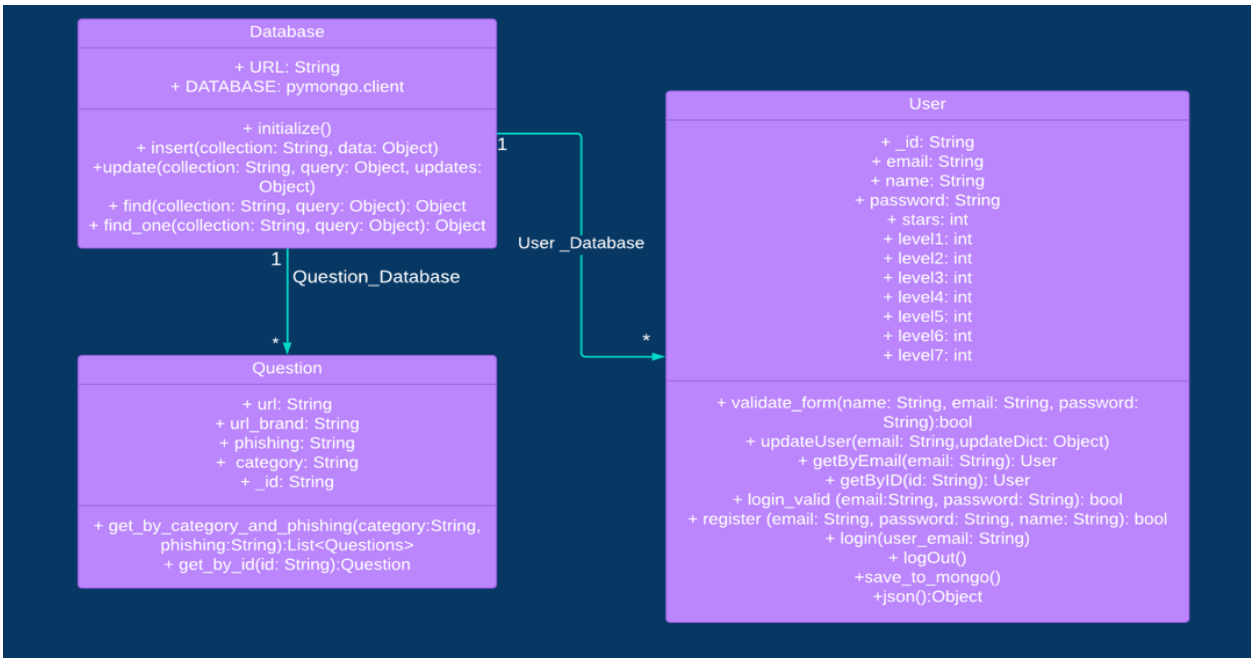

Fig3. Class Diagram of the Project

Action Flow of the Project

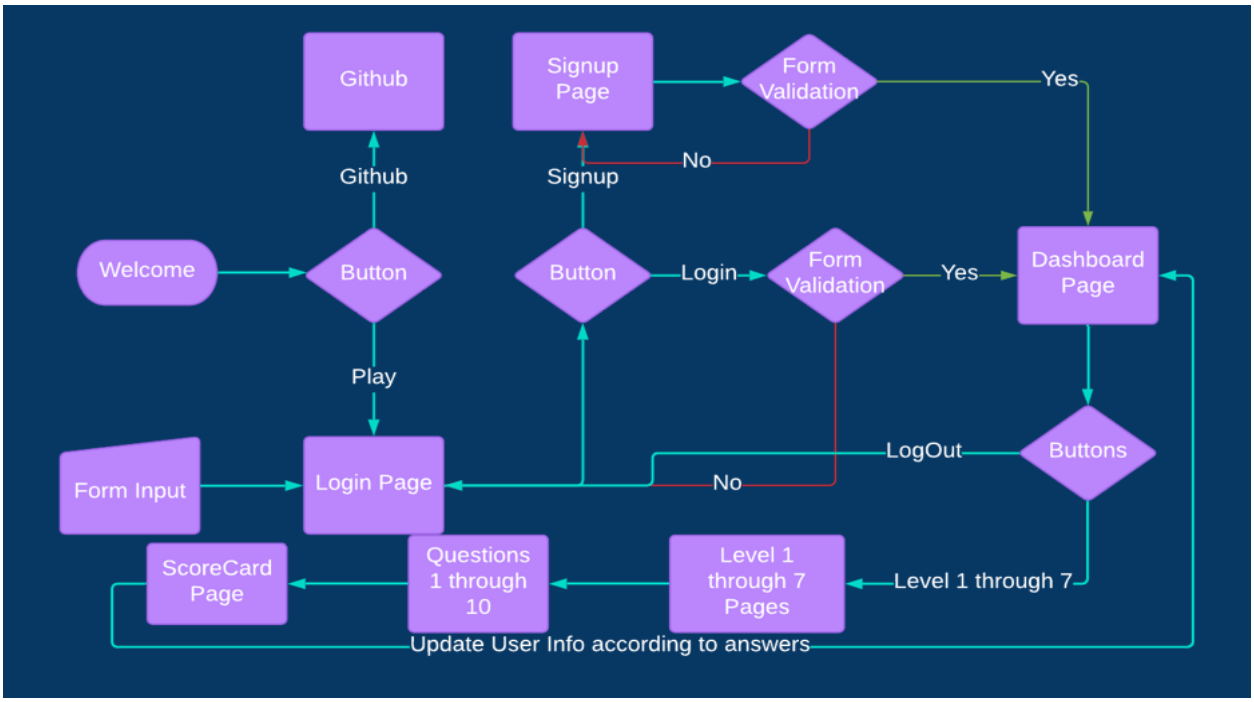

Fig4. Action Diagram of our WebApp

Datasets

We have used the following datasets to get the phishing and legitimate URLs,

1. **Aalto Phishstrom Dataset** [1] - This dataset contains a single CSV files which contains a tons of URLs with a lot of features which can be used with machine learning approaches for categorisation but for our use case we have only used **Domain** and **Label** attributes.
2. **UNB URL Dataset 2016** [2] - This dataset includes various types of emails - Benign URLs, Phishing URLs, Spam URLs, Malware URLs and Defacement URLs but we have only considered **Benign** and **Phishing URLs**. The data is stored in CSV files.

Categorization of URL Spoofing Tricks

Phishers apply several URL spoofing tricks. There are several approaches to categorize URL spoofing tricks. Different tricks have been explained in different levels of WhtThePhish. At each level of game, a spoofing trick is explained and a quiz is held for that spoofing trick. Correspondingly, we identified the following categories:

1. **IP Address URL without Brand:** In some cases, Phishers do not even bother registering any domain at all. In this spoofing trick, the host area of the URL contains an IP address while the path part does not contain the brand name,
   For example: http://5.178.64.164/secure to impersonate PayPal.
2. **Random/Unrelated/Trustworthy Domain, without Brand:** This trick uses random/unrelated or trustworthy names or strings as domain name and does not include the brand name of the targeted website in any other part of the URL.
   For example: http://www.szuhsa.fr/login.html, http://www.weather.com/ login.html or https://secure-payment.com to impersonate PayPal.

3. **Random/Unrelated/Trustworthy Domain, with Brand in Subdomain:** A phisher can include the brand name into the subdomain of a URL in combination with a random/unrelated/trustworthy domain name,
   For example: http://paypal.mark-chippy.com/account-setup/ or http://www.amazon.account.com/.

4. **Random/Unrelated/Trustworthy/IP Domain, with Brand in Path:** A phisher can include the brand name into the path part of a URL with a random/unrelated/trustworthy domain name.
   For example http://onlinepayment.com/www.paypal.com/.
   This attack can also happen in combination with an IP address URL.
   For example: http://5.178.64.164/paypal.

5. **Derivated Domains:** Sometimes, the modification of the original domain is registered by the phisher. In this case the modified domain contains the brand name in some form.
   For example: http://facebook-login.com can be registered in order to impersonate facebook.

6. **Introducing Typos:** In many cases, Phishers register domains which resemble the targeted domain but have typos, For example: the phisher can register micosoft.com to impersonate microsoft.com.
   One special case of introducing typo is swapping letters in the original domain name, e.g. http://mircosoft.com to impersonate Microsoft.

7. **Replacing Character(s):** Following this trick, A phisher can register domains where characters are replaced by other similar characters
   For example: https://www.arnazon.com.
   There are some more URL spoofing tricks which either cannot be recognized by the human eye (e.g. homograph attacks) or are irrelevant for our setting because they are redirected URLs such as tiny URLs or cloaked URLs.

Gamification

When the user opens the dashboard, there will be an option to select the levels. In total there are 7 levels. Initially only the first level will be unlocked for the user to play. To open the higher level, users need to play the lower level at least once. When the user opens the level, then a menu will appear where the user will get to know about the spoofing technique used in that level. After that users will have 10 yes/no questions. One can answer all of them or can skip a question. After the submit button is clicked, the next level gets unlocked and a page will be shown to the user where she/he will get their current score and maximum score. 7 levels are based on the 7 phishing techniques as shown in table 1.
The following game elements are used in our game to give users a good experience.

1. **Levels:** Leveling serves multiple purposes: First, it is important for the users to get a feeling for the progress they make. Second, it provides fixed points in the game from where they can restart or pause and continue the game later on. Finally, it enables you to increase the difficulty of the game with increasing levels.
2. **Achievements:** Achievements are special elements of a game that users can unlock if they, e.g. find a special object or if they play a certain level exceptionally well. This is in particular for people who are willing to invest a lot of time in a specific level in order to finish it perfectly or to find every hidden secret in it.
3. **User-Centered Design:** The design and implementation of a user-friendly and understandable app is achieved by giving extensive attention to the users' needs and wants as early as possible.

Learning Content per Part

The webapp entails the game with seven levels. Table 1 shows the link between the skills to properly judge on the trustworthiness of websites and the different parts of "WhtThePhish".

Taught Skill	Covered in
IP address, no brand	Level 1
Random/unrelated/trustworthy domain, no brand	Level 2
Random/unrelated/trustworthy domain, brand in subdomain	Level 3
Random/unrelated/trustworthy/IP domain, brand in path	Level 4
Derived domains	Level 5
Introducing typos	Level 6
Replacing character(s)	Level 7

Table 1: Spoofing Trick and Levels

Gallery for Our WebApp

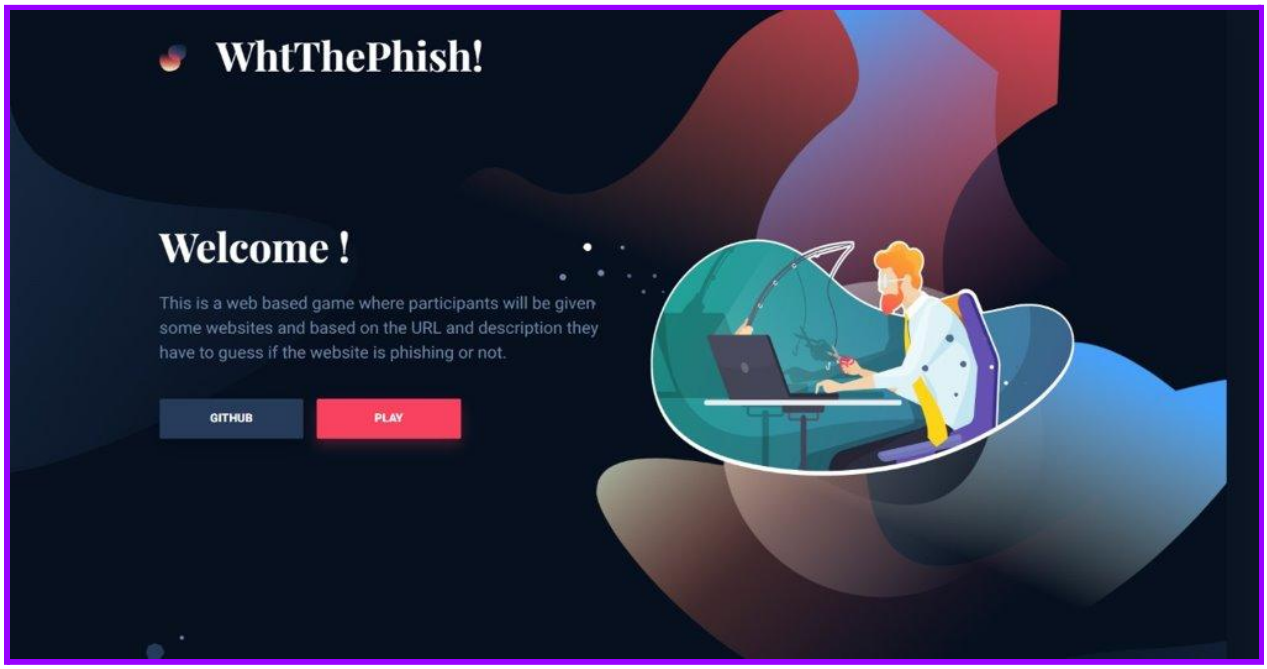

Fig. 5 Landing Page

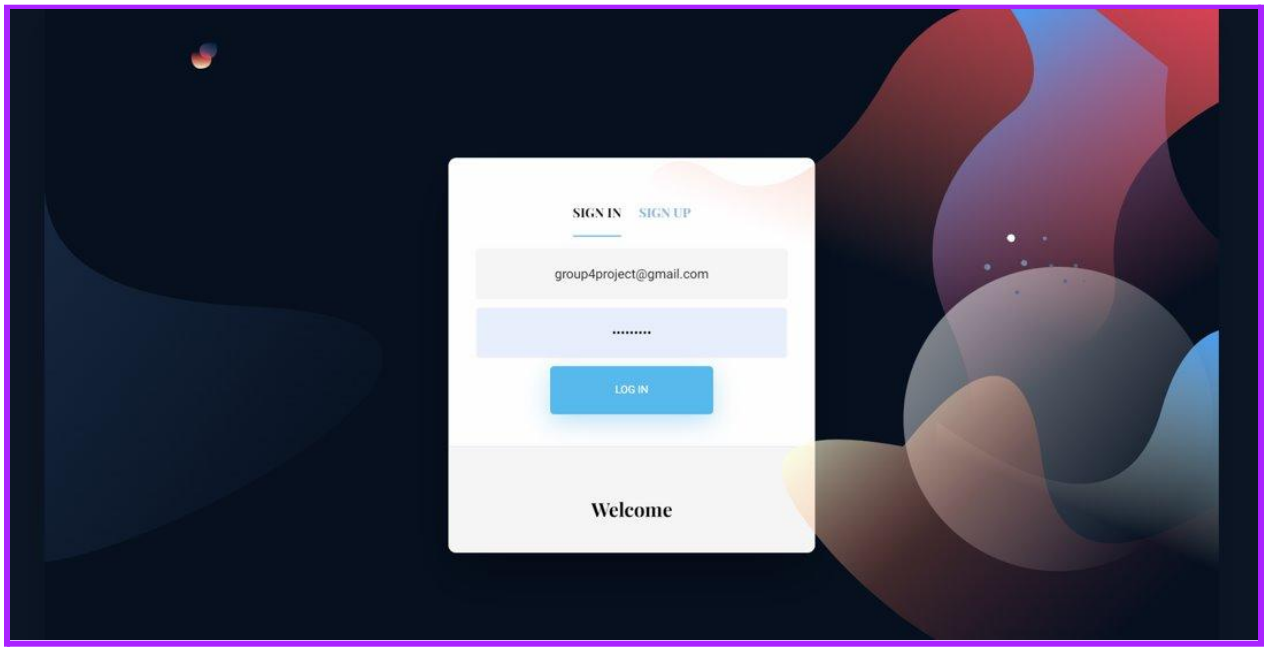

Fig. 6 Sign up/in page

Fig. 7. Game Dashboard

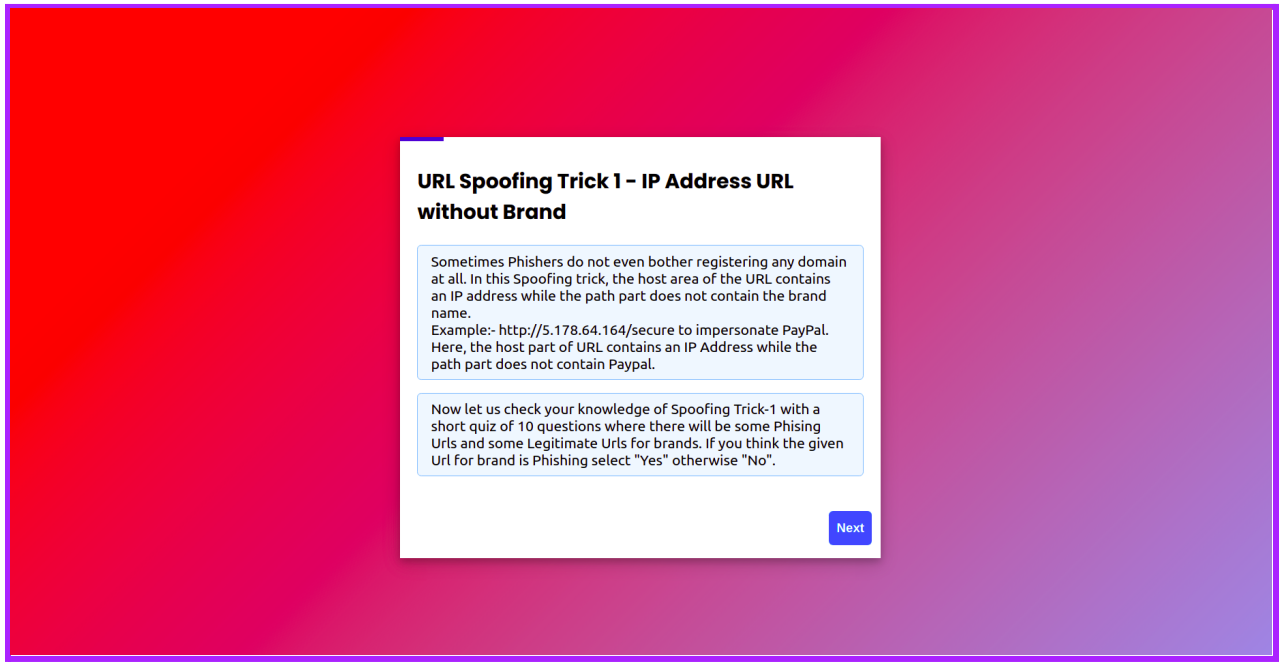

Fig. 8. Making user aware of the phishing technique

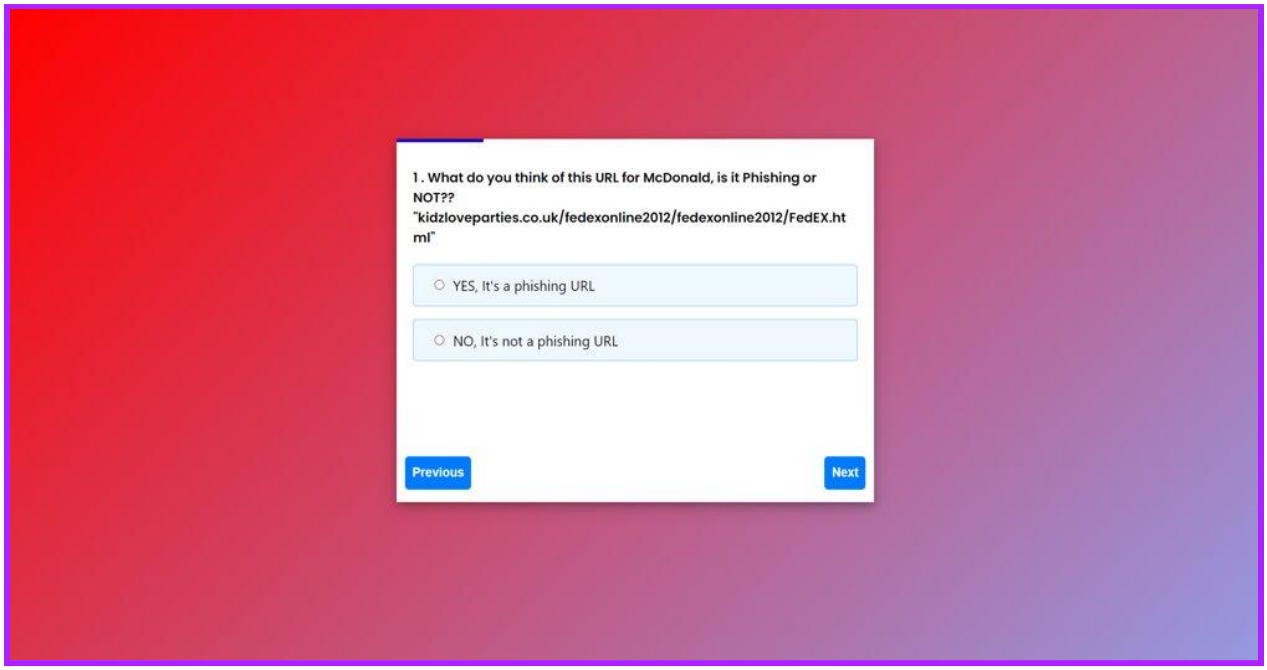

Fig. 9. Game Questions

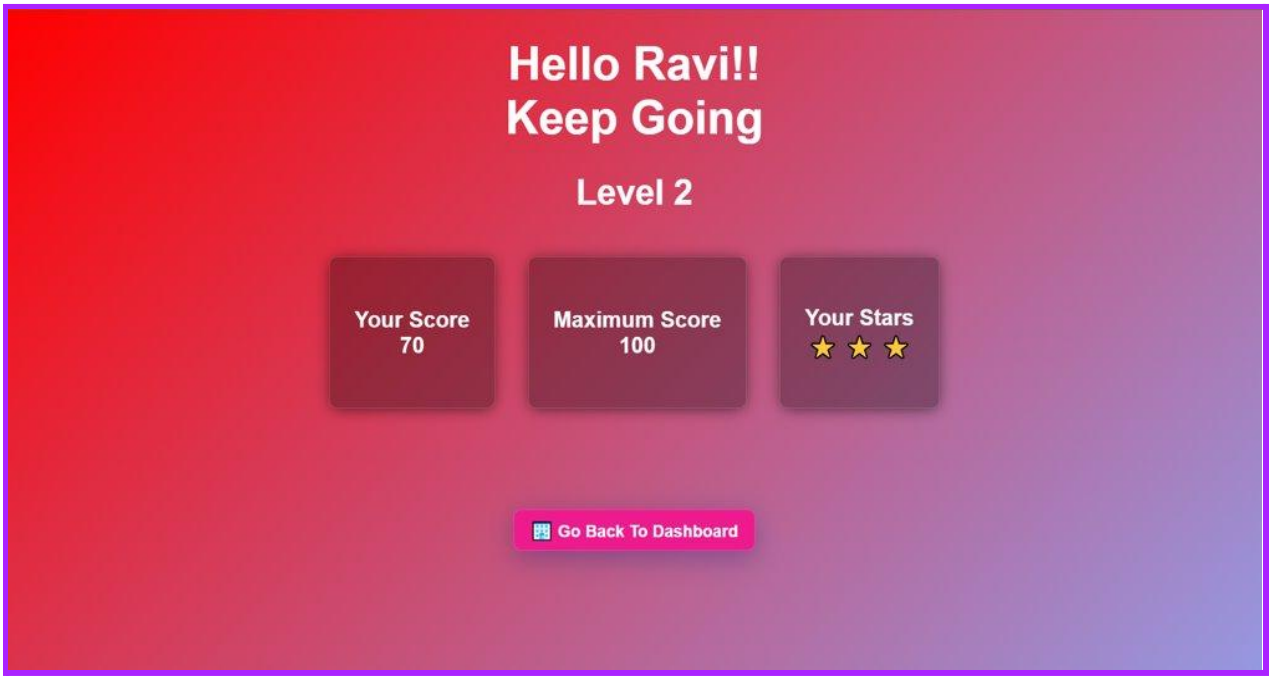

Fig. 10. Scorecard

Setting Up the Environment

1. Fork The repo
2. Clone it in your local machine using the forked repo
3. Create a virtual environment in your local machine in the folder you cloned the repo,

```
pip install virtualenv
virtualenv env
```

4. Activate the virtual env
   For Windows run

```
.\env\Scripts\activate
```

   For Linux run

```
source env/bin/activate
```

   Follow the instructions given after you create an env successfully .

5. Create a `secrets.py` file in the src folder with the following content.

```
USERNAME="[Your MongoDB Username]"
PASSWORD="[Your MONGODB Password]"
SECRET_KEY="[A Secret Key For the App]"
```

6. Then run

```
pip install -r requirements.txt
python -m src.common.insertQuestions
```

7. Now run the server
   For Windows run

```
python -m src.server
```

   For Linux run

```
python3 -m src.server
```

8. If you have reached this step, you are good to go.

Conclusion and Future Work

In the scope of this work, we have designed and implemented an anti-phishing education app – WhtThePhish – in a user-centered design approach. In a playful manner, users obtain valuable information on how to detect phishing URLs. The detection of phishing URLs is realized as a game, where the user can win points and after obtaining a certain threshold the user can move to the level ahead. We integrated learning principles and diverse gamification elements to engage the user in learning about phishing. In order to provide levels, with increasing difficulty to detect phishing URLs, we proposed a new categorization of URL spoofing tricks which we then explain in the different levels. Every starting page of a level consists of some information about that level and then the user can proceed to the quiz part. In future, we also plan to assess how such an education app can best be distributed. An idea would be to utilize embedded learning where simulated phishing emails are sent to users. Whenever users fall for such an email they could be prompted to download the education app. We have also thought about implementing a global leaderboard, reviewing each question as some of the future work.

References

[1] https://research.aalto.fi/en/datasets/phishstorm-phishing-legitimate-url-dataset

[2] https://www.unb.ca/cic/datasets/url-2016.html

[3] Canova, Gamze & Volkamer, Melanie & Bergmann, Clemens & Borza, Roland. (2014). NoPhish: An Anti-Phishing Education App. 188-192. 10.1007/978-3-319-11851-2_14.

[4] Sheng, Steve & Magnien, Bryant & Kumaraguru, Ponnurangam & Acquisti, Alessandro & Cranor, Lorrie & Hong, Jason & Nunge, Elizabeth. (2007). Anti-Phishing Phil: The design and evaluation of a game that teaches people not to fall for phish. ACM International Conference Proceeding Series. 229. 88-99. 10.1145/1280680.1280692.