STOR 565 - PROJECT PROPOSAL

Team Members: Egemen Elver, Malavika Mampally, Nghia Nguyen, Dilay Ozkan, Benjamin Torres

1. DATASET DESCRIPTION

Our research project utilizes the free and publicly available dataset of Hannouse and Yahiouche (2021), which includes 11430 URLs with 87 extracted features. The dataset also includes whether the provided URLs are legitimate, or they are created for phishing.

The features of the dataset are constructed based on three categories: URL-based, content-based, and external-based features, where our dataset has 56, 24, and 7 features for each category, respectively. The detailed description of each category is as follows (Hannouse and Yahiouche, 2021a):

- URL-based features are obtained by analyzing the text of the URLs. They contain structural and statistical features. Structural features are based on the position and presence of URL-based elements, such as the use of 'https' protocol. On the other hand, statistical-based features are related to the number of characters or specific words a URL contains. The number of dots or the length of the words can be considered as an example of statistical-based features.
- Content-based features are extracted by analyzing the HTML contents of the webpages of URLs. They are split into hyperlinks and abnormal content-based features. While the former is concerned with the number or the nature of the hyperlinks in HTML, the latter considers suspicious contents or behaviors such as empty links or different domain names.
- External-based features are constructed by querying the external services and search engines, such as page rank and domain age.

The response variable (label) of the dataset is binary as it has two classes: whether a URL is legitimate or not. Note that the dataset is balanced in terms of the labels, i.e. it contains an equal number of phishing and legitimate URLs.

2. MOTIVATION & GOALS

Have you ever received an email or text containing URLs from familiar or unfamiliar contacts? These could be sent for malicious purposes, such as what has become known as *phishing*. According to the European Union Agency for Cybersecurity (2024), phishing is a means to "persuade potential victims into divulging sensitive information such as credentials, or bank and credit card details." Utilizing a mixture of social engineering and deception, phishing attacks often occur over malicious webpages, e-mails, or instant messages that appear to be originating from a legitimate source. By pressure means in the form of scare tactics or urgent requests, the attackers try to induce the recipients to reply. The outgoing fraudulent messages tend not to be personalized and thus share many characteristics. Individuals or organizations with sufficient knowledge to identify a phishing website may be able to report or simply ignore the suspicious URL. Yet, even cyber experts are sometimes not able to identify a website with malicious intent. According to a study conducted by Intel, 97% of security experts fail to recognize phishing emails from genuine emails (Business Wire, 2015). Most people, particularly the elderly or vulnerable who lack the opportunities to obtain knowledge, are at risk. If they access URLs with a phishing intent, cybercriminals could steal their personal information, including but not limited to their SSN or banking details. Data from the Federal Bureau of Investigation's (FBI) Internet Crimes Report illustrated that in 2022, 300,497 phishing victims lost $52,089,159 just in the U.S. (FBI, 2022).

The origin of such monetary damages can be illustrated with the forthcoming example. You receive an email from an address that appears to be very similar to your bank's address, offering a $500 bonus just for logging in over the attached URL. When clicking on the URL, you see a website that is exactly the same as the official bank's webpage, except that the URL looks different. If you don't pay enough attention and log in, they might steal all the money in your account. More dangerously, thanks to the development of new technologies such as deep fake, cybercriminals can use this information to continue defrauding victims' relations by, for example, spreading texts or emails containing phishing URLs or malware, putting victims' relations in emergencies to ask for urgent money transfer, etc. Therefore, it is essential to develop some detectors to identify phishing web pages.

This work can benefit people in different ways. Of course, the application of this work needs the collaboration of experts from many different fields. As students in a Machine Learning class, we only consider two points of view: service providers/web browser developers and the authorities. From service providers/web browser developers' perspective, they can utilize these detectors to identify phishing web pages and warn users if they click on these URLs, or even block users from these phishing web pages. On the other hand, authorities, such as cyber police, can base their detection results on tracing where the cybercriminals place their servers and collaborating with local police to catch them. They can also require the hosting provider to close the phishing web pages.

Our goal is to develop binary classification models to categorize if a web page has malicious objectives, i.e. phishing, or if it is, in fact, legitimate. These models will be based on different machine learning methods, from traditional supervised methods, such as logistic regression, support vector machines, decision trees, and K-nearest neighbors, etc., to more sophisticated deep learning methods, such as among others, convolutional neural networks and long-short term memory. Based on this project, we can gain a deeper understanding of the methods we study in the class and compare their performance on a real dataset.

3. EXPLORATORY DATA ANALYSIS

For the data we are using, as mentioned before, we have variables from different categories: 56 variables come from the URLs of the web page, 24 variables are based on the content of the web page, and 7 variables are created using external sources. As the purpose is to identify phishing pages, a natural question that arises is, do we even have hope for this problem? Does the data seem to show some distinction between the two classes? While the aim of the project is to answer the former question, we can find a solution to the latter question by visualizing the 87 variables with dimension-reduction techniques.

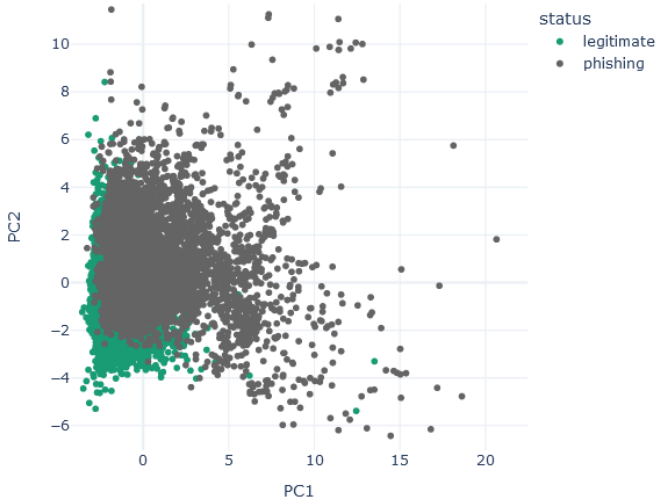

Figure 1: First two component of PCA, 15% variance explained.

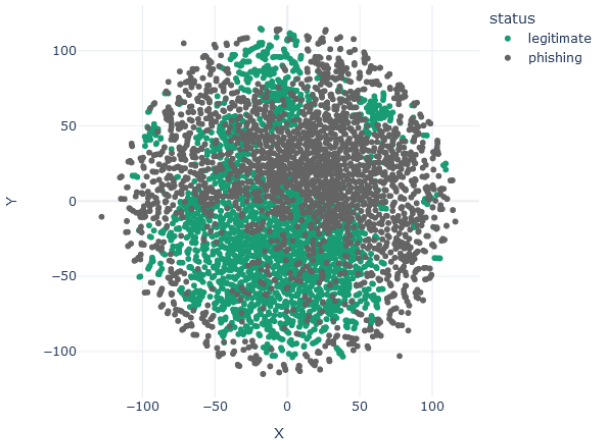

Figure 2: t-SNE with perplexity = 3

As we can see from Figure 1, PCA does not capture enough variability in the data as we are only able to explain 15% percent of it using the first two principal components. This could also be a reason why we do not have a clear distinction between legitimate and phishing URLs. It hints that the relationship may be complicated and non-linear. The cumulative variance plot of the PCA shows that more than 52 components are required to explain the variance of at least 90% as shown in Figure 3.

Figure 3: Cumulative variance plot.

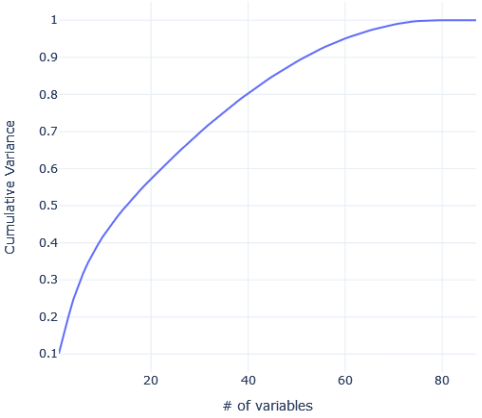

Figure 2, which was obtained by t-SNE method, gives us hope as it shows that the legitimate and phishing URLs might be distinguished from each other. Moreover, Figure 2 indicates there might be some clusters in the data.

Going on with exploring the data, the next natural question is, how related are the variables in the data set?

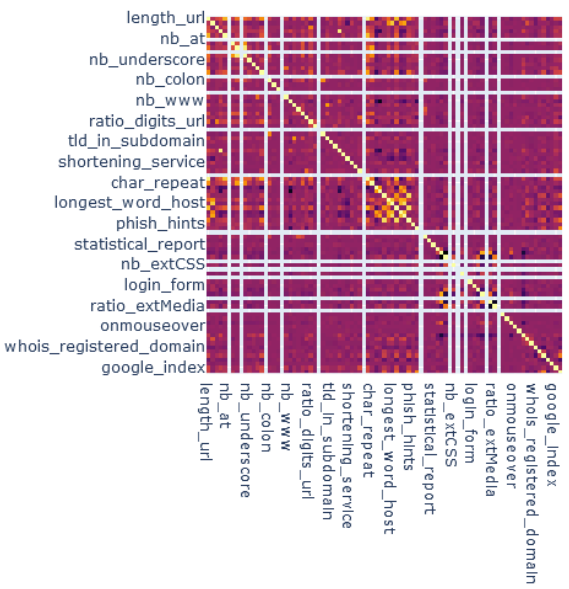

Figure 4: Correlation heatmap for legitimate URLs

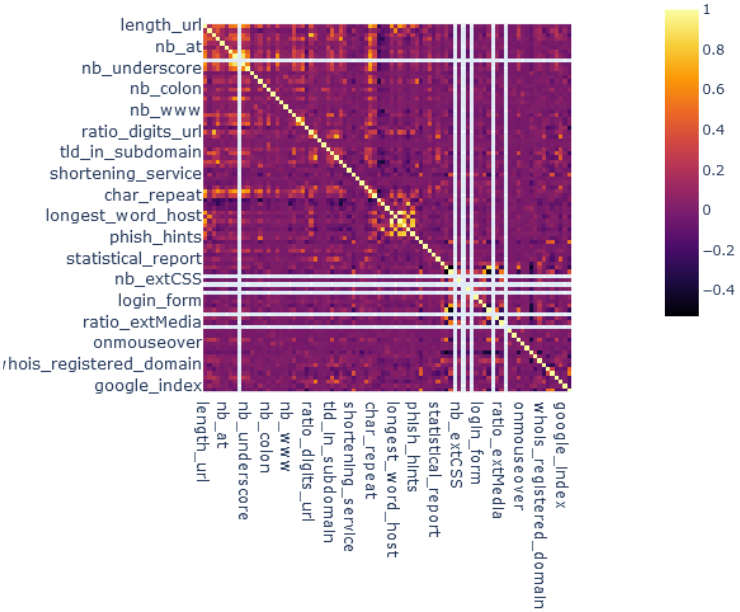

Figure 5: Correlation heatmap for phishing URLs

In Figure 4 and Figure 5, the white lines represent not-a-number values, this is due to constant columns (leading to zero variance). As we can see, legitimate web pages have more constant variables than phishing web pages. This might imply that some of the presence of the variables will help identify phishing web pages. We can see that the variables related to URLs are more related in the case of phishing web pages, this is because some phishing URLs may have strange-looking URLs as they have more dots, semicolons, and extra symbols.

Visualizing all 87 variables would not be effective; hence, we take a look at some specific variables to see any noteworthy pattern.

Let's examine the distribution of the URL lengths for each of the types of pages.

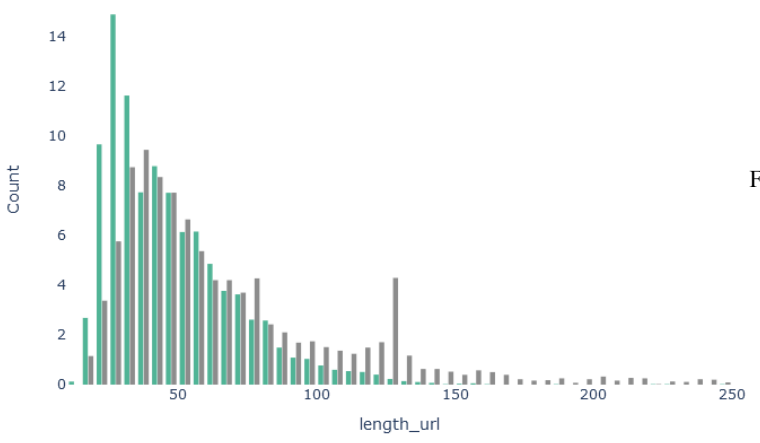

Figure 6: Histogram of URL lengths

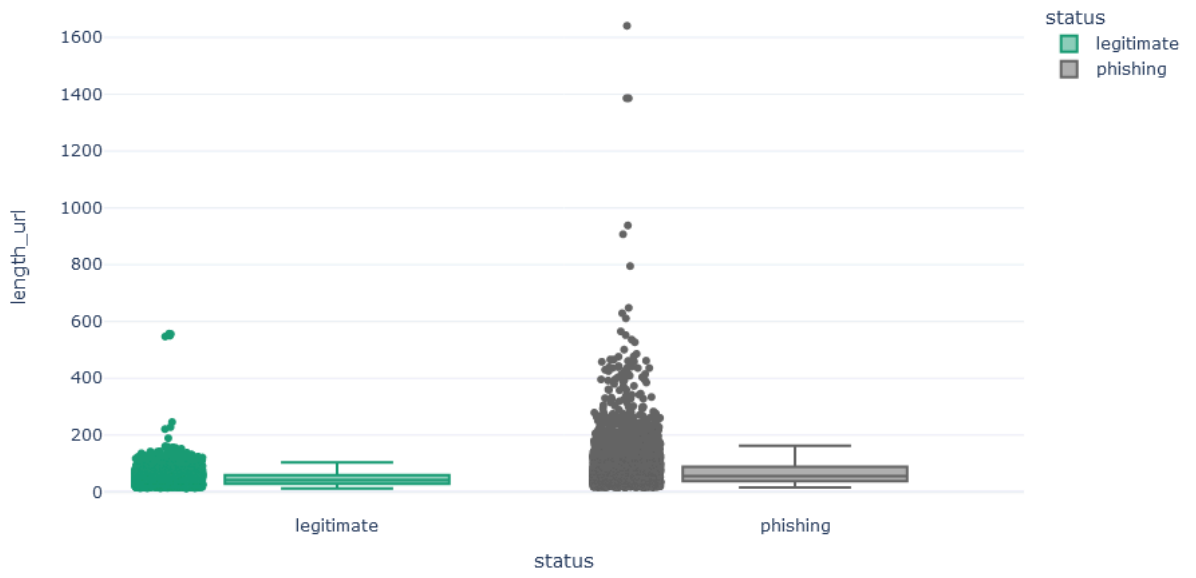

Figure 7: Boxplot of URL lengths

From the graphs provided in Figures 5 and 6, it can be observed that the legitimate URLs tend to be shorter than the phishing URLs. Such a finding is intuitive. Although a non-neglectable number of phishing URLs are short, there are cases of longer ones which will be a key to classify them. Although many would expect that a phishing URL can be identified due to its rather suspicious characteristics, such as uncommon letter combinations, this is not necessarily the case as we observe as follows.

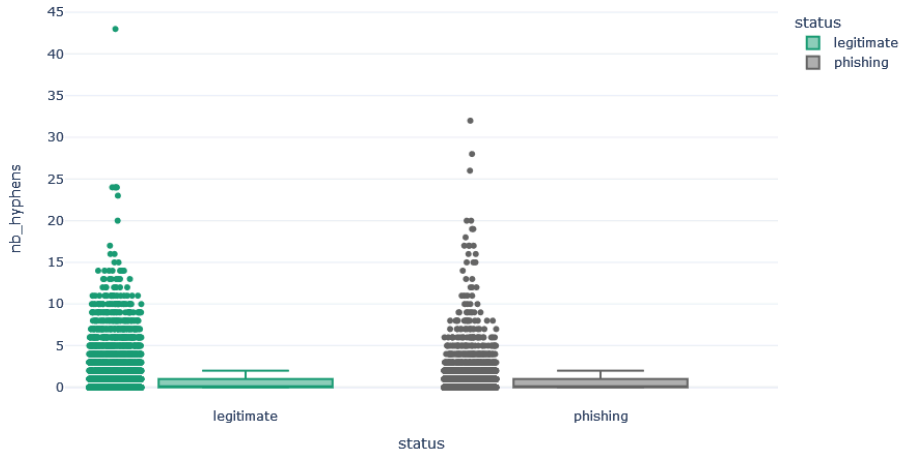

Figure 8: Box plot of the length of the number of hyphens in URLs

From the box plot, there is no clear distinction between the number of hyphens within each type of page. This hints that this variable may not be as useful by itself but could be useful when combined with other variables.

Another interesting thing we found by looking at Figure 8 is how many of the characters in the URL are digits. Well, here our intuition is correct, for phishing pages we do have a higher digit proportion, this could be because these pages are created automatically and hence named by a computer.

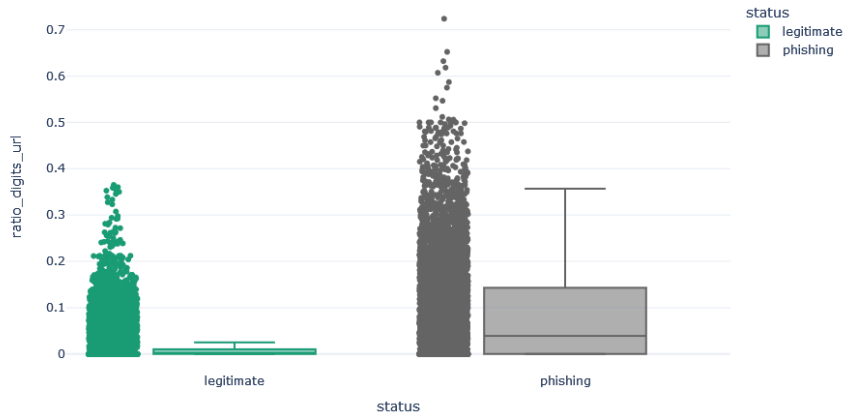

Figure 9: Box plot of the proportion of the digits in URLs

We found a contradicting insight about the number of hyperlinks in a URL. Based on an obvious assumption, we thought phishing pages would have more hyperlinks but in reality, legitimate links seemed to have more hyperlinks in general.

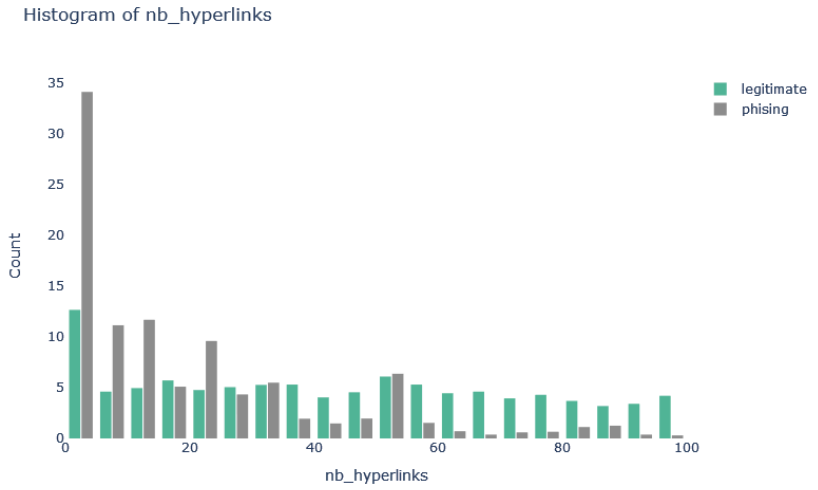

Figure 10: Histogram of the number of hyperlinks

Also, when we see the number of pop-up windows for each type of URL we see that there is not much distinction between legitimate URLs or phishing URLs. This is probably caused due to ads that are present in legitimate URLs.

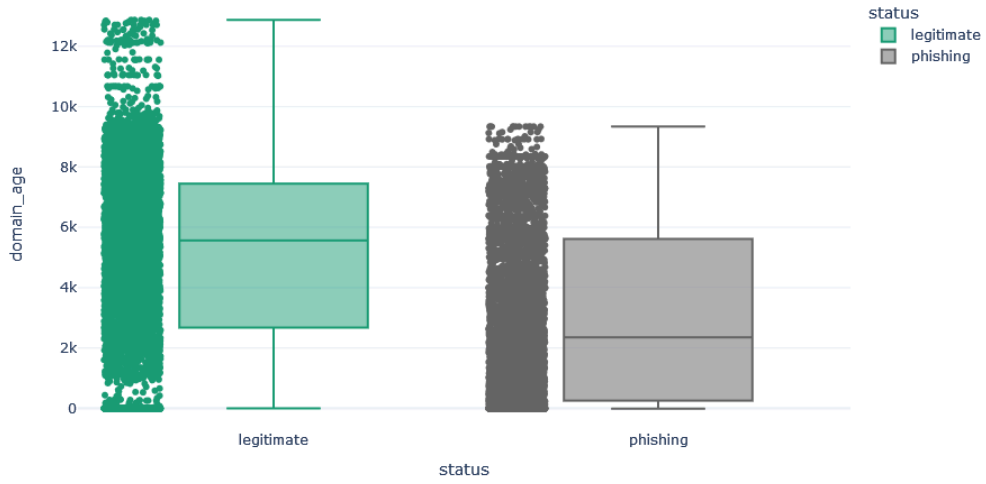

Figure 11: Box plot of the domain age

For the age of the URL, naturally, phishing URLs are newer than legitimate URLs. Even though we see some old phishing sites based on the plot, the majority of them are new. Lastly, we take a look at the web traffic in each type of URL.

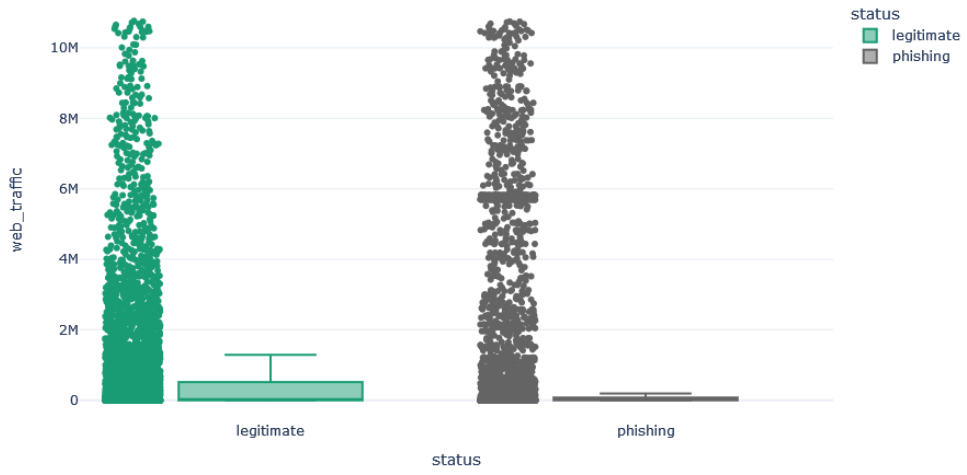

Figure 12: Box plot of the web traffic

For web traffic, we see a lower web traffic for phishing pages but we still have a good amount of pages with higher web flow which is surprising. Also, at around 6 million, we see an atypical behavior for phishing pages, a lot of pages have the same number which could be caused due to bots that enter the page in order to increase the rank of the page.

In summary, our exploratory data analysis reveals that phishing pages generally align with the expectations we normally had about them. These pages often feature large and unconventional characters, although not universally. Additionally, they may exhibit an increased user flow and a comparatively lower number of hyperlinks compared to legitimate pages.

REFERENCES

Business Wire. 2015. “97% Of People Globally Unable to Correctly Identify Phishing Emails.” May 12, 2015. https://www.businesswire.com/news/home/20150512005245/en/97-Of-People-Globally-Unable-to-Correctly-Identify-Phishing-Emails.

ENISA. n.d. “Phishing/Spear Phishing.” Page. ENISA. Accessed March 6, 2024. https://www.enisa.europa.eu/topics/incident-response/glossary/phishing-spear-phishing.

Federal Bureau of Investigation. 2022. “Internet Crime Report 2022.” https://www.ic3.gov/Media/PDF/AnnualReport/2022_IC3Report.pdf.

Hannousse, Abdelhakim, and Salima Yahiouche. 2021. “Web page phishing detection”, Mendeley Data, V3, doi: 10.17632/c2gw7fy2j4.3

Hannousse, Abdelhakim, and Salima Yahiouche. 2021a. “Towards Benchmark Datasets for Machine Learning Based Website Phishing Detection: An Experimental Study.” Engineering Applications of Artificial Intelligence 104 (September). https://doi.org/10.1016/j.engappai.2021.104347