STAT 423 Project Report - Phishing URL Analysis

Qirui Wang[1] and Yujia Dai[2]

Contributing authors: qw43@uw.edu; yd74@uw.edu;

1 Introduction

Phishing has become the mainstay of modern scams, and as more and more people start using the internet and electronics, more and more people are being targeted by such scammers. In this context, accurately recognizing phishing links and employing the necessary means to prevent more people from being scammed has become an important goal of cybersecurity. Some techniques that have been massively exploited are email filtering, anti-phishing tools, and URL inspection. One important part of these techniques is detecting phishing URLs. Thus, in this report, we mainly explore how to predict whether a link is a phishing link or not by URL link. In this report, we mainly explore how to predict whether a link is a phishing link or not by URL link.

The group members are Qirui Wang and Yujia Dai. Yujia Dai would be responsible for exploratory data analysis and feature engineering and Qirui would work on model training and evaluation. Since we only had 2 people, we worked on the report together, and both of us contributed much to the result and discussion. Here is our GitHub.

2 Data Description

The dataset we used is the Web page Phishing Detection Dataset[1] from Mendeley Data. The dataset includes 11034 URL observations with 87 features which are our predictor variables. The response variable is **status** which is the categorical variable with two levels: **legitimate** and **phishing**. We let 1 indicate the phishing website and 0 indicate the legitimate website. Features are from three different classes: 56 are extracted from the structure and syntax of URLs (URL-based features), 24 are extracted from the content of their correspondent pages (Content-based features), and 7 are extracted by querying external services (External-based feature). [2] Since not all URLs are accessible, we might ignore all Content-based features. The dataset is balanced in that 50% are phishing and 50 % are legitimate. We may not use all variables as our predictors, depending on our data cleaning of variables which will be discussed in the later section. All predictors before data cleaning are listed in the Appendix A.

URL-based features
URL-based features can be divided into *structural*, which are categorical variables, and *statistical* features, which are continuous variables. There are 16 structural-based features (IU1) that are concerned with the presence, position, and nature of URL-based elements. There are 40 Statistical-based features (IU2) concerned with the number or distribution of URL base elements, specific words, or characters in the text of URLs. Here are some URL-based features:

- URL parts lengths (IU2) *length_url*: Int, full URL length.
- IP (IU1) *ip*: 0/1, IP addresses are used in hostnames to hide the identity of websites. 1 indicates IP presents in hostnames, otherwise.
- Special Characters (IU2): Int, the number of occurrences of the following characters: “.”, “-”, “@”, ‘?”, “&”, “—”, “=”, “_”, “~”, “%”, “/”, “*”, “:”, “,”, “;”, “$”, “space”.

- HTTPS token (IU1) *https_token*: 0/1, most phishing websites do not provide any security facilities compared with legitimate ones, so the use of HTTPS is a legitimacy indicator. 1 indicates the use of HTTPS, 0 otherwise.

External-based features
External features (E) are obtained by querying reference third-party services and search engines. Here are some external-based features:

- Domain age *domain_age*: int, represents the age of URL domains. Usually, phishing websites are short-lived.
- Google index *google_index*: 0/1, web pages not indexed by Google are supposed phishing. 1 represents not indexed by Google, 0 otherwise.

3 Research Questions

This is a list of research questions that are to be addressed during the project:

- Can we find a linear classifier to classify legitimate URLs and phishing URLs?
- Is a URL with a lot of special characters necessarily a phishing URL?
- Which URL-based features are more important in phishing URL classification? Are External-based features useful in phishing URL classification?

4 Methodology

Since we want to make a prediction about a categorical variable **status**, we would choose **logistic regression** to assess the association between variables.

4.1 Exploratory Data Analysis

First, we check whether there are missing values in our dataset, the result shows that the dataset doesn't contain any missing value. Then, by looking through our dataset, we notice that some observations contain negative values, which is impossible. So we remove those observations that contain negative values in any columns, and then we finally have 9586 observations.

We count the categorical predictors to check the balance of each level of variables. We observed that *statistical_report* has three levels (0,1,2) but it should only have two levels (0,1) stated in the original paper. So we remove this variable. From Table 1 shown below, we notice that many categorical variables are highly unbalanced. So we decided to disregard these unbalanced variables and keep *ip*, *https_token*, *prefix_suffix*, and *google_index* as our categorical predictors. There is a mosaic plot for these four variables in Appendix B which would be helpful to see the distribution of these four categorical predictors and the response variable.

Predictor	0	1	Predictor	0	1
ip	8165	1421	**https_token**	3800	5786
punycode	9584	2	port	9564	22
tld_in_path	9035	551	tld_in_subdomain	9125	461
abnormal_subdomain	9455	131	**prefix_suffix**	7601	1985
random_domain	8934	652	shortening_service	8265	1321
path_extension	9584	2	domain_in_brand	8405	1181
brand_in_subdomain	9549	37	brand_in_path	9555	31
suspecious_tld	9469	117	whois_registered_domain	9342	244
dns_record	9475	111	**google_index**	4484	5102

Table 1: Table to count each level of categorical variables

We found that the summary statistics (min, max, median, and mean) for our predictor variables *nb_or* and *nb_external_redirection* are all 0. This means that for all the observations we looked at, these two variables always had a value of 0. So, we chose to take these predictors out of our analysis.

Predictor	min	max	median	mean
nb_or	0	0	0	0
nb_external_redirection	0	0	0	0

Table 2: Summary of predictor

To check multicollinearity in regression, we first use a full model (47 predictors) to find the variance inflation factor (VIF). We keep variables with a VIF below 10 and remove those with a VIF above 10 from our data. We combine all special characters into a single column named *special_char*. After that, we end up with 21 variables to use for the model selection.

We create the correlation matrix for the 21 predictors below. We notice that *ratio_digits_url* and *ip* are highly correlated, so are *page_rank* and *domain_age*, and we should be aware of such problem.

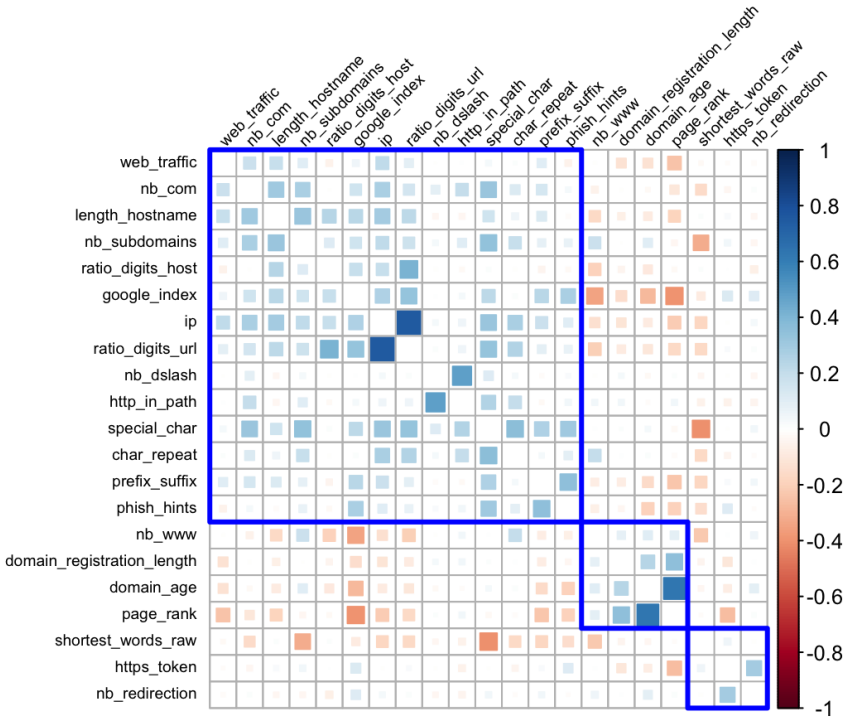

Fig. 1: Correlation between each predictor

We get the summary statistics for our 17 continuous variables among 21 predictors as follows. We notice that the mean and median of *domain_registration_length* and *web_traffic* have large differences, which indicates the variables are right-skewed, so we may perform the log transformation to make it closely follow the normal distribution. Other predictors would roughly follow the normal distribution.

Predictor	min	max	median	mean
length_hostnam	6	214	19	21.74494
special_char	4	111	9	9.54705
nb_www	0	2	0	0.44951
nb_com	0	5	0	0.13447
nb_dslash	0	1	0	0.00532
http_in_path	0	4	0	0.01221
ratio_digits_url	0	0.72	0	0.05192
ratio_digits_host	0	0.64	0	0.01944
nb_subdomains	0	3	2	2.22981
nb_redirection	0	5	0	0.49750

char_repeat	0	146	3	2.97100
shortest_words_raw	1	31	3	3.12821
phish_hints	0	7	0	0.31139
domain_registration_length	0	29829	270	543.01325
domain_age	0	12874	5075	4839.88661
web_traffic	0	10767986	1737	849076.73065
page_rank	0	10	3	3.38713

Table 3: Summary of continuous variables

The Fig.2 support our thought that *domain_registration_length* and *web_traffic* are right-skewed. The histogram for *log(web_traffic)* and *log(domain_registration_length)* looks much better and follows the normal distribution. Also *length_hostname* and *domain_age* seem to be normally distributed.

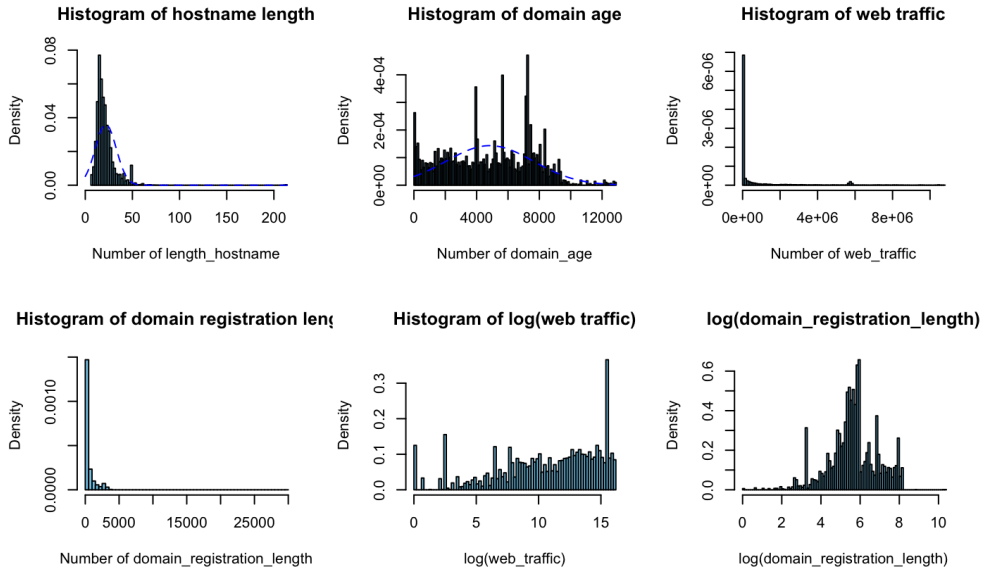

Fig. 2: Histogram of predictors

4.2 Model Selection

With the cleaned dataset, we split the data into training and test with 90% of the data points in the training subset and the remaining 10% in the test subset. Then we would select the best model using AIC and Mallows's Cp. Because our training dataset contains 8681 observations which is a large number of observations after data cleaning and traintest dataset split. Besides, we would test the model performance on test data, thus we would not choose BIC and Kfold CV since their model choice may have somewhat worse predictive performance.

We perform forward and backward model selection using AIC and Mallows's Cp criteria. In the AIC selection process, the selected models we got from backward and forward selection are the same with the AIC score of 3552. Using Mallows's CP selection process, the selected models we got from backward and forward selection are also the same with Mallows's Cp value of 18.89373. Finally, we got the model with the following predictor variables: length_hostname, ip, special_char, nb_www, nb_com, nb_dslash, http_in_path, https_token, ratio_digits_url, ratio_digits_host, nb_subdomains, nb_redirection, char_repeat, phish_hints, domain_registration_length, domain_age, web_traffic, google_index, page_rank.

5 Results and Discussion

5.1 Model Summary

The model selected by the AIC has 18 predictors, while the model selected Mallows's Cp has 19 predictors, including *char_repeat*. The following tables are the summary for both models. We notice that the p-value of *domain_registration_length* is larger than 0.05, which means this predictor is not significant. This is also supported by the confidence interval table (Table 6) listed later.

Variable	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	1.1283658	0.2442534	4.620	3.84e-06 ***
google_index1	3.2012509	0.1005954	31.823	$< 2e-16$ ***
page_rank	-0.4746114	0.0253086	-18.753	$< 2e-16$ ***
nb_www	-2.0784721	0.1041997	-19.947	$< 2e-16$ ***
phish_hints	2.0347914	0.1187782	17.131	$< 2e-16$ ***
web_traffic	-0.0994396	0.0073993	-13.439	$< 2e-16$ ***
ratio_digits_host	8.7224748	1.2160413	7.173	7.35e-13 ***
ip1	1.4578743	0.2214668	6.583	4.62e-11 ***
domain_age	-0.0002096	0.0000217	-9.661	$< 2e-16$ ***
nb_dslash	4.8076270	0.9011905	5.335	9.57e-08 ***
nb_redirection	-0.3078317	0.0631644	-4.874	1.10e-06 ***
length_hostname	0.0353352	0.0065909	5.361	8.27e-08 ***
https_token1	-0.7069716	0.1035346	-6.828	8.59e-12 ***
special_char	-0.0833830	0.0119502	-6.978	3.00e-12 ***
nb_subdomains	0.3809410	0.0882407	4.317	1.58e-05 ***
http_in_path	1.1425648	0.4310410	2.651	0.00803 **
ratio_digits_url	2.7571842	1.0090160	2.733	0.00628 **
nb_com	0.3759345	0.1506509	2.495	0.01258 *
domain_registration_length	0.0527708	0.0280076	1.884	0.05954

Table 4: Logistic Regression Coefficients selected by AIC

Variable	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	1.132e+00	2.444e-01	4.632	3.61e-06 ***
length_hostname	3.546e-02	6.589e-03	5.381	7.42e-08 ***
ip1	1.407e+00	2.364e-01	5.951	2.66e-09 ***
special_char	-8.448e-02	1.217e-02	-6.941	3.89e-12 ***
nb_www	-2.090e+00	1.060e-01	-19.723	$< 2e-16$ ***
nb_com	3.799e-01	1.508e-01	2.520	0.01175 *
nb_dslash	4.817e+00	9.013e-01	5.345	9.03e-08 ***
http_in_path	1.137e+00	4.318e-01	2.634	0.00843 **
https_token1	-7.029e-01	1.037e-01	-6.776	1.23e-11 ***
ratio_digits_url	2.850e+00	1.022e+00	2.789	0.00528 **
ratio_digits_host	8.671e+00	1.219e+00	7.116	1.11e-12 ***
nb_subdomains	3.743e-01	8.891e-02	4.209	2.56e-05 ***
nb_redirection	-3.063e-01	6.321e-02	-4.846	1.26e-06 ***
char_repeat	6.645e-03	1.084e-02	0.613	0.53987
phish_hints	2.035e+00	1.189e-01	17.111	$< 2e-16$ ***
domain_registration_length	5.275e-02	2.801e-02	1.883	0.05964 .
domain_age	-2.099e-04	2.171e-05	-9.669	$< 2e-16$ ***
web_traffic	-9.923e-02	7.407e-03	-13.396	$< 2e-16$ ***
google_index1	3.203e+00	1.007e-01	31.820	$< 2e-16$***
page_rank	-4.741e-01	2.533e-02	-18.720	$< 2e-16$ ***

Table 5: Logistic Regression Coefficients selected by Cp

5.2 Comparing Model

Because the two selected models are different and one includes the other. Thus, we would use χ^2 test for nested model comparison. The result is:

```
Analysis of Deviance Table

Model 1: status ~ google_index + page_rank + nb_www + phish_hints + web_traffic +
    ratio_digits_host + ip + domain_age + nb_dslash + nb_redirection +
    length_hostname + https_token + special_char + nb_subdomains +
    http_in_path + ratio_digits_url + nb_com + domain_registration_length
Model 2: status ~ length_hostname + ip + special_char + nb_www + nb_com +
    nb_dslash + http_in_path + https_token + ratio_digits_url +
    ratio_digits_host + nb_subdomains + nb_redirection + char_repeat +
    phish_hints + domain_registration_length + domain_age + web_traffic +
    google_index + page_rank
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      8662     3513.9
2      8661     3513.5  1  0.38453   0.5352
```

Since the p-value is 0.5352 which is large, we would choose the previous model as our final model - the model selected by AIC with 18 predictors.

5.3 Final Model

The following is our logistic regression model that was selected by AIC.

$$\begin{aligned}\log(\frac{P(\text{status}=1)}{1-P(\text{status}=1)} &= 1.1283658 + 3.2012509 \times \text{google\_index1} - 0.4746114 \times \text{page\_rank} \\ &- 2.0784721 \times \text{nb\_www} + 2.0347914 \times \text{phish\_hints} - 0.0994396 \times \text{web\_traffic} \\ &+ 8.7224748 \times \text{ratio\_digits\_host} + 1.4578743 \times \text{ip1} - 0.0002096 \times \text{domain\_age} \\ &+ 4.8076270 \times \text{nb\_dslash} - 0.3078317 \times \text{nb\_redirection} \\ &+ 0.0353352 \times \text{length\_hostname} - 0.7069716 \times \text{https\_token1} \\ &- 0.0833830 \times \text{special\_char} + 0.3809410 \times \text{nb\_subdomains} \\ &+ 1.1425648 \times \text{http\_in\_path} + 2.7571842 \times \text{ratio\_digits\_url} \\ &+ 0.3759345 \times \text{nb\_com} + 0.0527708 \times \text{domain\_registration\_length}\end{aligned}$$

According to the model summary shown in section 5.1, we found that only *domain_registration_length* has a pvalue greater than 0.05 which means this predictor variable is not significant at 5% level. All other predictor variables are at least significant at the 5% level.

5.4 Residual Plots

From the first plot, we can see that there is a horizontal line which means that there is a constant variance. From the scalelocation plot, there is roughly a horizontal line which means the assumptions of equal variance are met. From the residuals vs leverage plot, we find although some points are close to the Cook's distance line, they don't fall outside of the dashed line. This means there are not any influential points in our regression model.

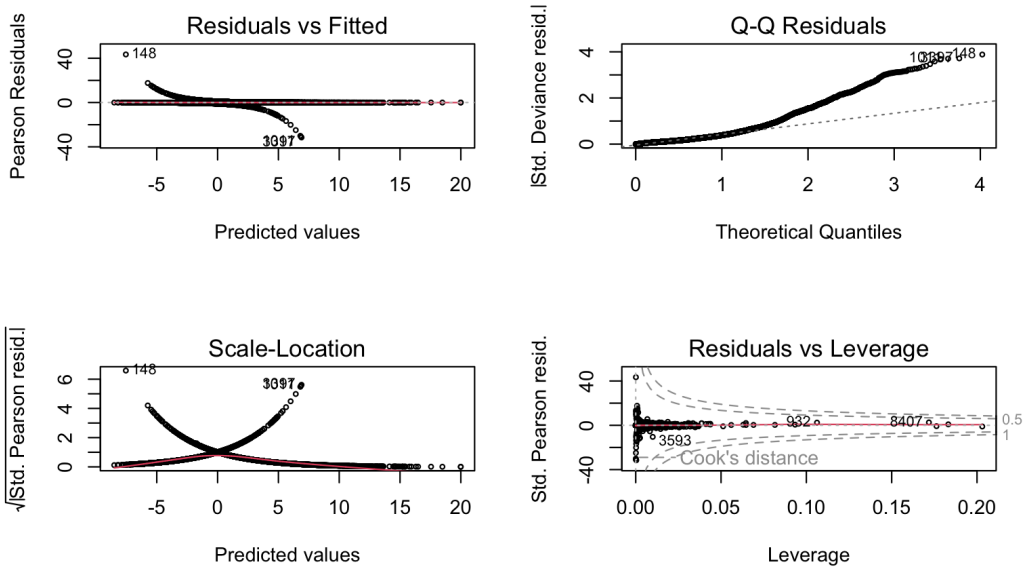

Fig. 3: Residual Plots

5.5 Confidence Interval

We further calculated the confidence interval for each predictor variable. Here is the confidence interval for our predictors. Based on the table, we notice the confidence interval for *domain_registration_length* contains 0, which indicates this predictor is not significant.

Variable	**2.5 %**	**97.5 %**
(Intercept)	0.6504	1.6082
google_index1	3.0063	3.4007
page_rank	-0.5247	-0.4255
nb_www	-2.2847	-1.8761
phish_hints	1.8057	2.2714
web_traffic	-0.1140	-0.0850
ratio_digits_host	6.3798	11.1469
ip1	1.0240	1.8925
domain_age	-0.0002524	-0.0001674
nb_dslash	3.1111	6.6909
nb_redirection	-0.4320	-0.1842
length_hostname	0.0225	0.0484
https_token1	-0.9111	-0.5051
special_char	-0.1076	-0.0607
nb_subdomains	0.2084	0.5544
http_in_path	0.2906	1.9852
ratio_digits_url	0.7944	4.7508
nb_com	0.0821	0.6723
domain_registration_length	-0.0019	0.1079

Table 6: Confidence Interval

5.6 Performance on Test Dataset

We test the model's predictive performance on the test dataset which contains 905 observations and the accuracy is $834/905 \approx 0.92$.

6 Conclusion

Based on our analysis, we would say we could find a linear classifier to classify legitimate URLs and phishing URLs. Since we fit our model with the training dataset and test the accuracy by the test dataset, we get a good accuracy of about 92%. The special character in our model turned out to be very important for guessing if a URL is phishing or not because it has a very small p-value, showing it's a key factor in making predictions. However, even with this information, we can't conclude that a URL with a lot of special characters is always a phishing site.

From Table 4, we could see which variables are significant in our phishing URL classification. In **URL-based features**, *nb_www*, *phish_hints*, *ratio_digits_host*, *ip1*, *nb_dslash*, *nb_redirection*, *length_hostname*, *https_token1*, *special_char*, *nb_subdomains*, *http_in_path*, *ratio_digits_url*, and *nb_com* are important. In **External-based features**, *google_index1*, *page_rank*, *domain_age*, and *web_traffic* are important.

However, our analysis is only based on logistic regression and the result might not be able to generalize to all situations in most recent online phishing URLs. Thus, more works and edge cases should be discussed in the future. More data should be collected from different types of phishing URLs and more advanced anomaly detection should be done to obtain a higher predictive performance.

Appendix A All predictors variable description [2]

If not specified otherwise, its data type is Int.
URL-based features

- length_url, **length_hostname**: Full URL length, full hostname length
- **ip**: 0/1, IP addresses are used in hostnames to hide the identity of websites. IPs can also be used without dots or hexadecimal encoded. The presence of IPs in any format in hostnames is considered as phishing indicator.
- **secial_char** (nb_dots, nb_hyphens, nb_at, nb_qm, nb_and, nb_eq, nb_underscore, nb_tilde, nb_percent, nb_slash, nb_star, nb_colon, nb_comma, nb_semicolumn, nb_dollar, nb_space): the number of occurrences of the following characters.
- common terms (**nb_www**, **nb_com**, **nb_dslash**, **http_in_path**): Common terms in URLs such as 'www', '.com', 'http' and '//' are used only once in legitimate URLs where it is observed that they are used more than once in phishing URLs.
- **https_token**: 0/1, most phishing websites do not provide any security facilities compared with legitimate ones. Thus, the use of HTTPS is a legitimacy indicator.
- Ratio of digits (**ratio_digits_url**, **ratio_digits_host**): Float, a high number of digits in URLs is considered as a phishing indicator. We consider the ratio of digits in full URLs and hostnames.
- punycode: 0/1, punycode is used in domain names to replace some ASCIIs with Unicode characters. URLs with punycodes are considered phishing.
- port: 0/1, port numbers are rarely used in legitimate URLs. Therefore, URLs with the port indicator are considered phishing.
- TLD position (tld_in_path, tld_in_subdomain): 0/1, top-level domains (TLDs) appear only before the path. When TLDs appear in the path or in the subdomain part, the URL is considered phishing.
- abnormal_subdomain: 0/1, Phishing URLs may use the following pattern 'w[w]?[0-9]*' instead of 'www' to deceive users. URLs with subdomains matching such patterns are considered phishing.
- **nb_subdomains**: Phishing URLs use more subdomains compared with legitimate ones. Thus, the number of subdomains is a phishing feature.
- **prefix_suffix**: 0/1, when"-" is found in domain names, the URL is considered phishing.
- random_domain: 0/1, phishing URLs use words formed from random characters. Domain names are checked for randomness.
- shortening_service: 0/1, URL shortening service is used to indicate short URLs that serve as a redirect to other long and complex URLs. The use of a shortening service is considered a phishing indicator.
- path_extension: 0/1, Malicious scripts can be added to legitimate pages. Some file extensions used in URL paths may launch such kinds of attacks. The presence of the following malicious path extensions is considered: 'txt', 'exe', 'js'.
- Redirections (**nb_redirection**, nb_external_redirection): URL redirection is a technique used to open pages with different URLs than those initially selected by users. The number of redirections and external redirections are considered phishing indicators.
- NLP features (length_words_raw, **char_repeat**, **shortest_words_raw**, shortest_word_host, shortest_word_path, longest_words_raw, longest_word_host, longest_word_path, avg_words_raw, avg_word_host, avg_word_path): Natural language processing and word-raw features are also used in phishing detection. We consider the number of words, char repeat, the shortest words in URLs, hostnames, paths, and the longest words in URLs, hostnames, paths, and the average length of words in URLs, hostnames, and paths.
- **phish_hints**: Phishing URLs use sensitive words to gain trust on visited web pages. The number of such words in URLs is considered as phishing indicator.
- Brand domains (domain_in_brand, brand_in_subdomain, brand_in_path): 0/1, Phishing URLs use brand domain names in different URL parts. The presence of brand names in the domain part is considered as a legitimacy indicator where their presence in subdomains or paths is considered as a phishing indicator.
- suspecious_tld: 0/1, TLDs are checked for suspiciousness.
- statistical_report: 0/1, URL domains are checked if their IP addresses match one of the top phishing domains.

External-based features

- whois_registered_domain:0/1, domains of phishing websites do not match any WHOIS database record contrary to most legitimate domains. Therefore, URLs with domains not registered in WHOIS are considered phishing.
- **domain_registration_length**: Phishing websites live for a short time, while legitimate websites are regularly paid for several years in advance. Instead of proposing a specific threshold as proposed in [20, 18, 27], we use the number of years the domain renewal amount was paid as a phishing indicator.
- **domain_age**: Since phishing websites are short-lived, the age of URL domains is considered a phishing indicator.
- **web_traffic**: Phishing websites generally have less number of visitors compared with legitimate websites.
- dns_record: 0/1, Domain Name Server (DNS) is mandatory to retrieve the IP address of URLs for access. Therefore, URL domains must be registered within the DNS. A missing DNS record is a phishing indicator.
- **google_index**: 0/1, Phishing websites live for short times and are often accessible through direct links sent to users in emails, they do not need to be indexed by Google. Web pages not indexed by Google are supposed phishing.
- **page_rank**: Phishing web pages are not very popular, hence, they are supposed to have low page ranks compared with legitimate web pages.

Appendix B Mosaic plot

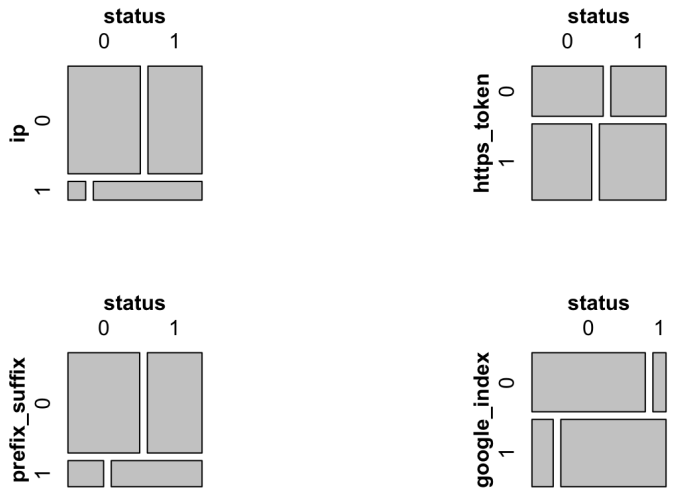

Fig. B1: Mosaic plot

References

[1] Hannousse, A., Yahiouche, S.: Web page phishing detection. https://data.mendeley.com/datasets/c2gw7fy2j4/3 (2021)

[2] Hannousse, A., Yahiouche, S.: Towards benchmark datasets for machine learning based website phishing detection: An experimental study. Engineering Applications of Artificial Intelligence **104**, 104347 (2021) https://doi.org/10.1016/j.engappai.2021.104347