Statistical Inference

https://github.com/monastyrskyy/bootcamp_repo/tree/master/Capstone_1_Phishing

The barplot visualization shows that some features stray further from the sample proportion of phishing websites than others. To develop some hypotheses about what features will be influential in model selection, I first filtered down to the 'interesting' features with the following method. For each category of each feature, I ran a 10,000 size bootstrap sample of the 'result' label variable. I then took this information to calculate a 95% confidence interval for the proportion that phishing websites made up of that group. If the overall proportion of phishing websites in the whole sample was outside of the 95% CI for any of the categories of a feature, I labeled that feature as 'interesting'. The number of phishing websites that have these 'interesting' features will either be significantly higher or lower than the sample average, at 95% confidence. This might then suggest that these features will have a lot of influence when building models and may have high predictive power in whether a website is phishing or legitimate. Below is a list of the 'interesting' columns.

```
url_contains_double_slash
favicon
standard_ports
url_contains_https_token
submitting_to_email
redirect
on_mouseover
right_click_disabled
popup_window
uses_iframe
```

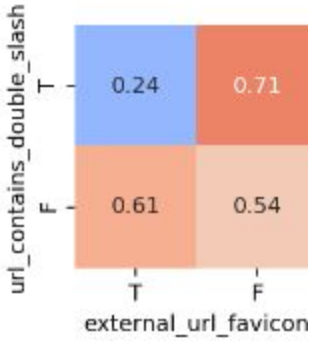
Merely looking at a list of feature names doesn't tell me too much, so the next step would be to visualize the features somehow.

Exploratory Data Analysis - Closer Look

To visualize the data within the 'interesting' columns, I want to display the proportion of phishing websites within each combination of categories for each combination of features. Below is a crosstab example of what I mean.

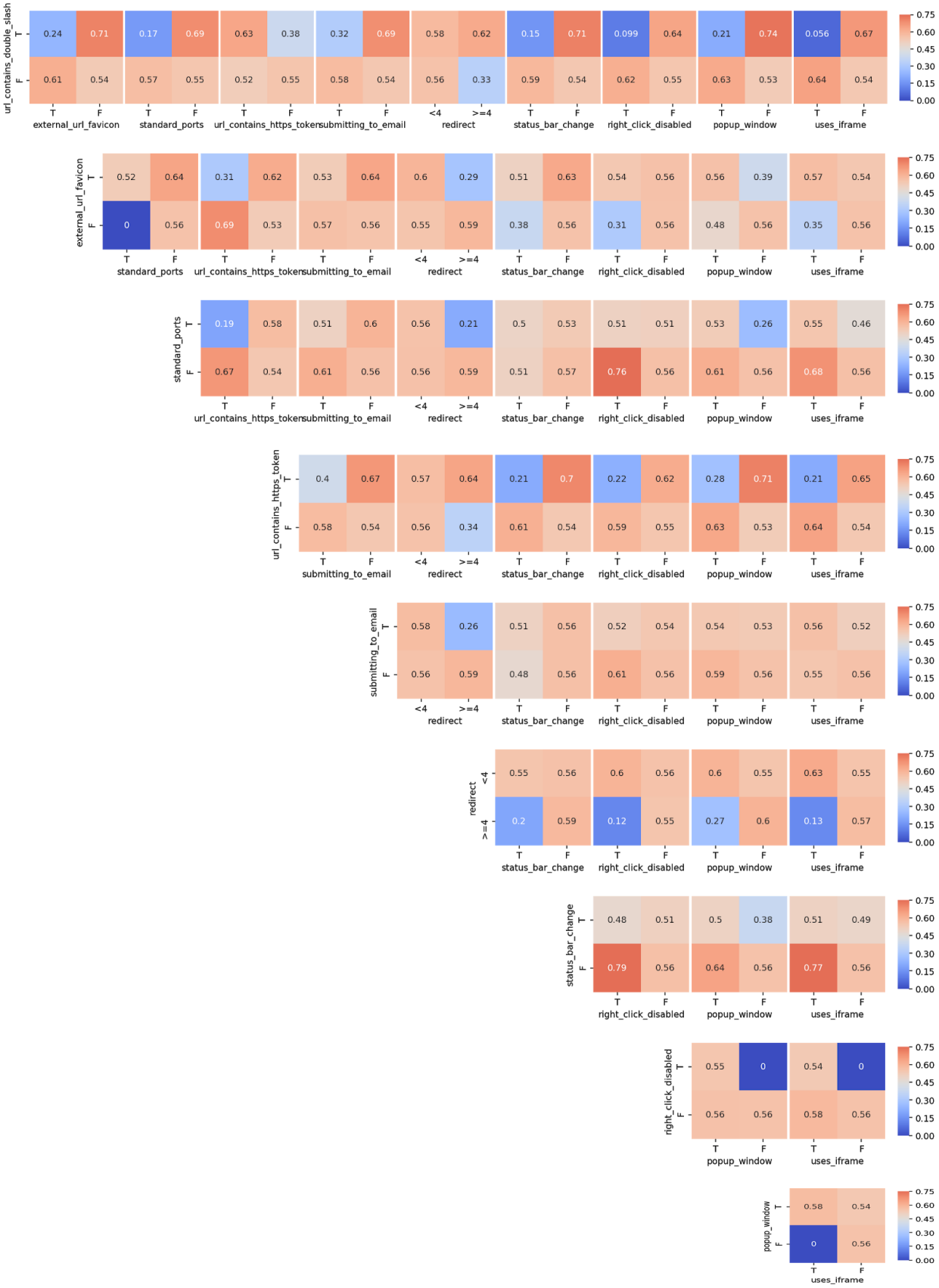
external_url_favicon	-1	1
url_contains_double_slash		
-1	0.240506	0.710692
1	0.614853	0.535176

In this crosstab, each combination of each category is plotted and the number it contains inside is the proportion of phishing websites of that group. Notice that this crosstab only compares two features: 'external_url_favicon' to 'url_contains_double_slash', whereas in my full analysis I want to compare each combination of features, and visually display the proportions of phishing websites.



The above graph is a visualization of the previously discussed crosstab. The color switches from shades of blue to shades of red, depending on whether the proportion of phishing websites for each specific combination in the visualization is above or below the sample proportion of phishing websites. The darker the color, the bigger the absolute difference between the combination proportion and the sample proportion.

To get the above information for each combination of features, I iterated over each combination of features and created a dictionary of such crosstab data frames as the one above but for every combination of features. Below is a visualization of each combination in a heatmap format. Notice that the combinations with dark colors are easy to see, and the trends are easy to see. For example among the websites that contain a double slash in the url, only 21% of those that also contain a popup window are phishing, as opposed to the 74% of those that don't contain a popup. This example shows that there are a lot of non-intuitive but interesting insights that this visualization can provide.


url_contains_double_slash
T 0.24 0.71 0.17 0.69 0.63 0.38 0.32 0.69 0.58 0.62 0.15 0.71 0.099 0.64 0.21 0.74 0.056 0.67
F 0.61 0.54 0.57 0.55 0.52 0.55 0.58 0.54 0.56 0.33 0.59 0.54 0.62 0.55 0.63 0.53 0.64 0.54
T F T F T F T F <4 >=4 T F T F T F T F T F
external_url_favicon standard_ports url_contains_https_token submitting_to_email redirect status_bar_change right_click_disabled popup_window uses_iframe
external_url_favicon
T 0.52 0.64 0.31 0.62 0.53 0.64 0.6 0.29 0.51 0.63 0.54 0.56 0.56 0.39 0.57 0.54
F 0 0.56 0.69 0.53 0.57 0.56 0.55 0.59 0.38 0.56 0.31 0.56 0.48 0.56 0.35 0.56
T F T F T F <4 >=4 T F T F T F T F T F
standard_ports url_contains_https_token submitting_to_email redirect status_bar_change right_click_disabled popup_window uses_iframe
standard_ports
T 0.19 0.58 0.51 0.6 0.56 0.21 0.5 0.53 0.51 0.51 0.53 0.26 0.55 0.46
F 0.67 0.54 0.61 0.56 0.56 0.59 0.51 0.57 0.76 0.56 0.61 0.56 0.68 0.56
T F T F <4 >=4 T F T F T F T F T F
url_contains_https_token submitting_to_email redirect status_bar_change right_click_disabled popup_window uses_iframe
url_contains_https_token
T 0.4 0.67 0.57 0.64 0.21 0.7 0.22 0.62 0.28 0.71 0.21 0.65
F 0.58 0.54 0.56 0.34 0.61 0.54 0.59 0.55 0.63 0.53 0.64 0.54
T F <4 >=4 T F T F T F T F T F
submitting_to_email redirect status_bar_change right_click_disabled popup_window uses_iframe
submitting_to_email
T 0.58 0.26 0.51 0.56 0.52 0.54 0.54 0.53 0.56 0.52
F 0.56 0.59 0.48 0.56 0.61 0.56 0.59 0.56 0.55 0.56
<4 >=4 T F T F T F T F T F
redirect status_bar_change right_click_disabled popup_window uses_iframe
redirect
<4 0.55 0.56 0.6 0.56 0.6 0.55 0.63 0.55
>=4 0.2 0.59 0.12 0.55 0.27 0.6 0.13 0.57
T F T F T F T F
status_bar_change right_click_disabled popup_window uses_iframe
status_bar_change
T 0.48 0.51 0.5 0.38 0.51 0.49
F 0.79 0.56 0.64 0.56 0.77 0.56
T F T F T F
right_click_disabled popup_window uses_iframe
right_click_disabled
T 0.55 0 0.54 0
F 0.56 0.56 0.58 0.56
T F T F
popup_window uses_iframe
popup_window
T 0.58 0.54
F 0 0.56
T F
uses_iframe
0.75 0.60 0.45 0.30 0.15 0.00