Phishing Websites

I am using a '**TreeBag**' model to learn from the features in the training dataset. Here, we have an extensive set of features available in the dataset, ranging from simple attributes like 'Domain Length' to more complex ones like 'Port being used', etc.

I was able to get an accuracy of **0.964** with a treeBag model from **Caret** package. The dataset is having approximately **2500** instances of **30** features, out of which some **1900** instances are used for training purposes. This is quite a good ratio for using a tree based bagging model. With enough instances we can be pretty sure that model won't overfit the training data and as apparent from the results it doesn't, resulting in an accuracy of more than **96%**.

An ensemble tree based model works really good in this scenario and after having a look at the variable importances of different attributes we can conclude along with with our real world experience that the most important variables found by the treeBag model are indeed the most significant ones while identifying a phishing website. Here are 5 most important attributes as found by,

R Script (treebag):

1. (Abnormal URL Anchor)url_of_anchor: 100.000
2. (SSL State) ssl_state: 97.690
3. (Traffic) traffic: 76.097
4. (Prefix Suffix) pref_suf: 62.365
5. (Domain Age) domain_Age: 9.584

BigML:

1. (SSL State) ssl_state: 35.64%
2. (Abnormal URL Anchor) url_of_anchor 14.82%
3. (Prefix Suffix) pref_suf : 12.54%
4. (URL has subdomain) has_sub_domain: 6.01%
5. (Domain Age) domain_Age: 5.09%

Although there is a slight variation in the most important variables in RScript and BigML, but that of course it attributes to the slightly different implementations of bagging method in **Caret** and **BigML.**