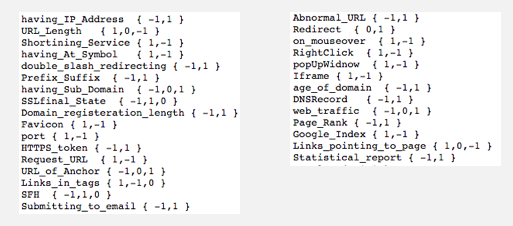
Rod Ta

CSC 59929

Classifying Phishing Websites

My project is on Classifying Phishing websites. The dataset includes 30 features shown below.



I applied four different machine learning methods to the dataset trying to figure out what method would be the best fit for my dataset. The four methods I used were Naïve Bayes, Logistic Regression, Decision Tree, and SVM (Support Vector Machine).

Naïve Bayes:

There were three different types of Naïve Bayes: Bernoulli, Gaussian and Multinomial.

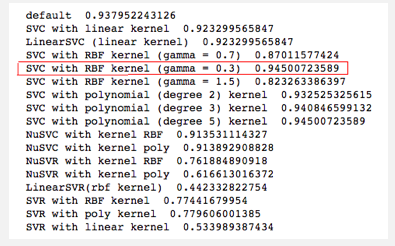
Bernoulli gave me the best result of 90.3%

Multinomial gave me 77.5%

Gaussian gave me 60.3%

The second method I used was Logistic Regression. Using Logistic Regression was able to give me an accuracy of 92.6%

The third method I used was Decision Tree. Decision Tree was able to give me one of the best accuracy of 95.3%

The last method I used was SVM. I experimented with many different SVMs which all gave me very different results: 

The best result was SVC with RBF kernel with gamma = 0.3, which gave me an accuracy of 94.5%

SVCs were all able to give me very good accuracy, while SVRs gave me very poor accuracy. The results made sense since SVRs are used for continous datasets, while SVC are used for classification.

Upon doing futhur research, out of the 30 features, the 9 most important features are in the order of: Prefix\_suffix, Url of Anchor, SSL\_State, Has\_sub\_Domain, Traffic, Req\_URL, Long\_domain, Domain\_Age and Tag\_Links.

I then used the Decision Tree Classifiers on these 9 features to see how well it will do in predicting phishing websites, and it gave me a result of 94.4% accuracy, which is really good.

In my project, I also found out that having more features does not necessary mean you will always get a better accuracy. This is very apparent when I used the Gaussian Naïve Bayes classification. Please refer to the charts I included at the end of the report for visualization. This only goes to prove that choosing the right features is very important to go along with the right model.

In addition, I also used these four machine learning methods on different various amount of features and displayed their accuracy along with the average accuracy

The bar charts showed the accuracy of using only 1 features all the way to using all 30 features at once.

