Capstone Project: Website Phishing Data Set

Introduction

Phishing is a form of identity theft that is usually made through emails or website in order to gain authorized access to user’s private information. Phishing websites are bogus sites, where a phishing attacker attracts victims to a spoofed website similar to a legitimate one, a so-called “phishing site”. Once victims access the phishing site, then the attacker attempts to convince them to send their private information such as usernames, passwords and credit card resulting in stealing their information that might cost them over a billion dollars each year.

Data

We use the Phishing Websites dataset available at Machine Learning Repository from UCI website. The dataset consists of 11055 collected Websites samples. 4898 of them are labeled as phising sites while the remaining 6157 samples are labeled as legitimate. The choice of this dataset is due to its richness in extracted features from various categories (it has 30 features divided into four groups) as we will describe in the next subsection.

Link : <http://archive.ics.uci.edu/ml/datasets/Website+Phishing>

As discussed the features were classified into below four categories.

### Address Bar based Features---12 features

### Abnormal Based Features---6 features

### HTML and JavaScript based Features---5 features

### Domain based Features---8 features

### Address Bar based Features

### Using the IP Address

### Long URL to Hide the Suspicious Part

### Using URL Shortening Services “TinyURL”

### URL’s having “@” Symbol

### Adding Prefix or Suffix Separated by (-) to the Domain

### Sub Domain and Multi Sub Domains

### HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)

### Domain Registration Length

### Favicon

### Using Non-Standard Port

### The Existence of “HTTPS” Token in the Domain Part of the URL

### Abnormal Based Features

### Request URL

### URL of Anchor

### Links in <Meta>, <Script> and <Link> tags

### Server Form Handler (SFH)

### Submitting Information to Email

### Abnormal URL

### HTML and JavaScript based Features

### Website Forwarding

### Status Bar Customization

### Disabling Right Click

### Using Pop-up Window

### IFrame Redirection

### Domain based Features

### Age of Domain

### DNS Record

### Website Traffic

### PageRank

### Google Index

### Number of Links Pointing to Page

### Statistical-Reports Based Feature

And throughout this project below were the naming conventions used.

* " 1 " ----> Legitimate
* " -1 "---> Phishing
* " 0 "----> Suspicious

DATA WRANGLING

Using dplyr, the Database was imported and the required information was read using read\_CSV. The database contains information about features of the website that were collected and hence there were no missing values in any of the columns. Furthermore, to analyze about the features separately, we subset them into 4 different categories as mentioned above. Also, a website ID column is added to the dataframe.

#### EXPLORATORY ANALYSIS

To begin the analysis of finding the highly important features bar charts are plotted . This would give us fair idea which feature plays a prominent role in phishing website detection.

In these Bar charts, each feature is plotted separately for phishing websites and legitimate websites. So for these bar plots we can say how many legitimate websites were predicted correctly.

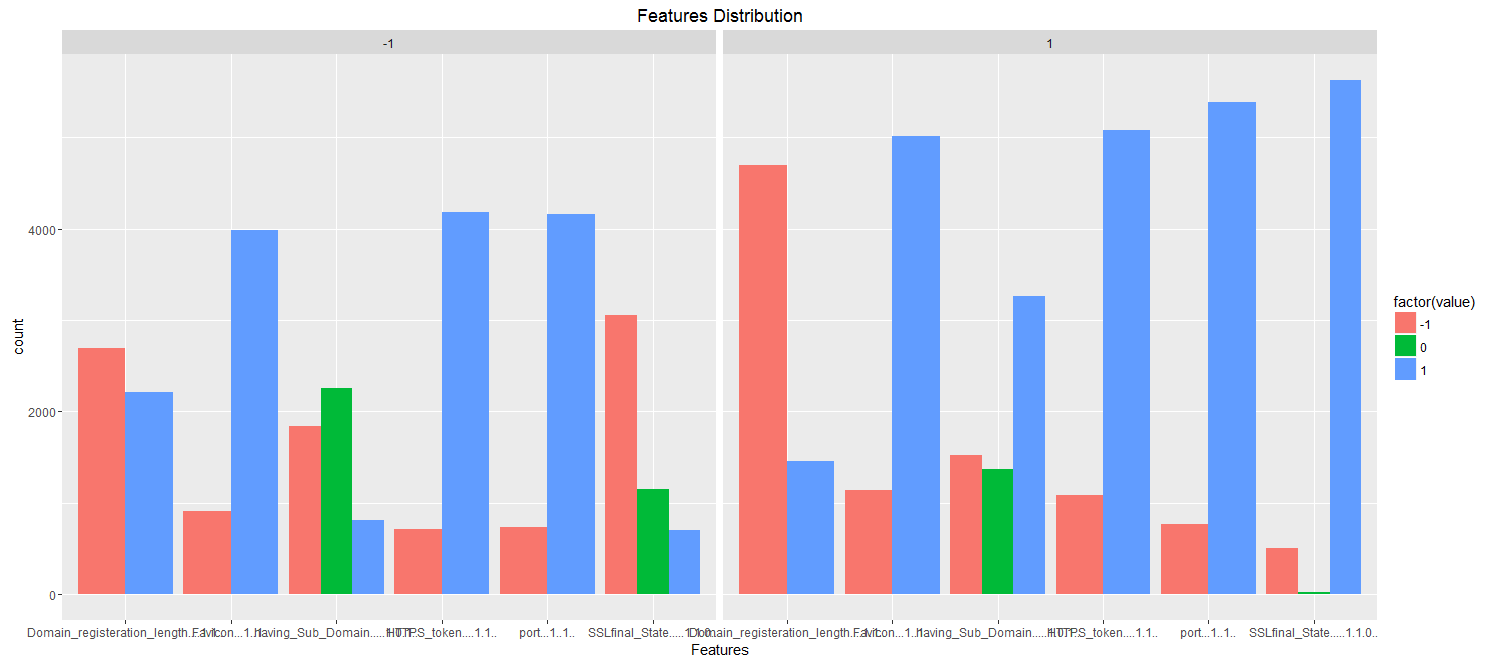
From the below bar chats we can say Two features play very important role in predicting the website is legitimate or phishing.

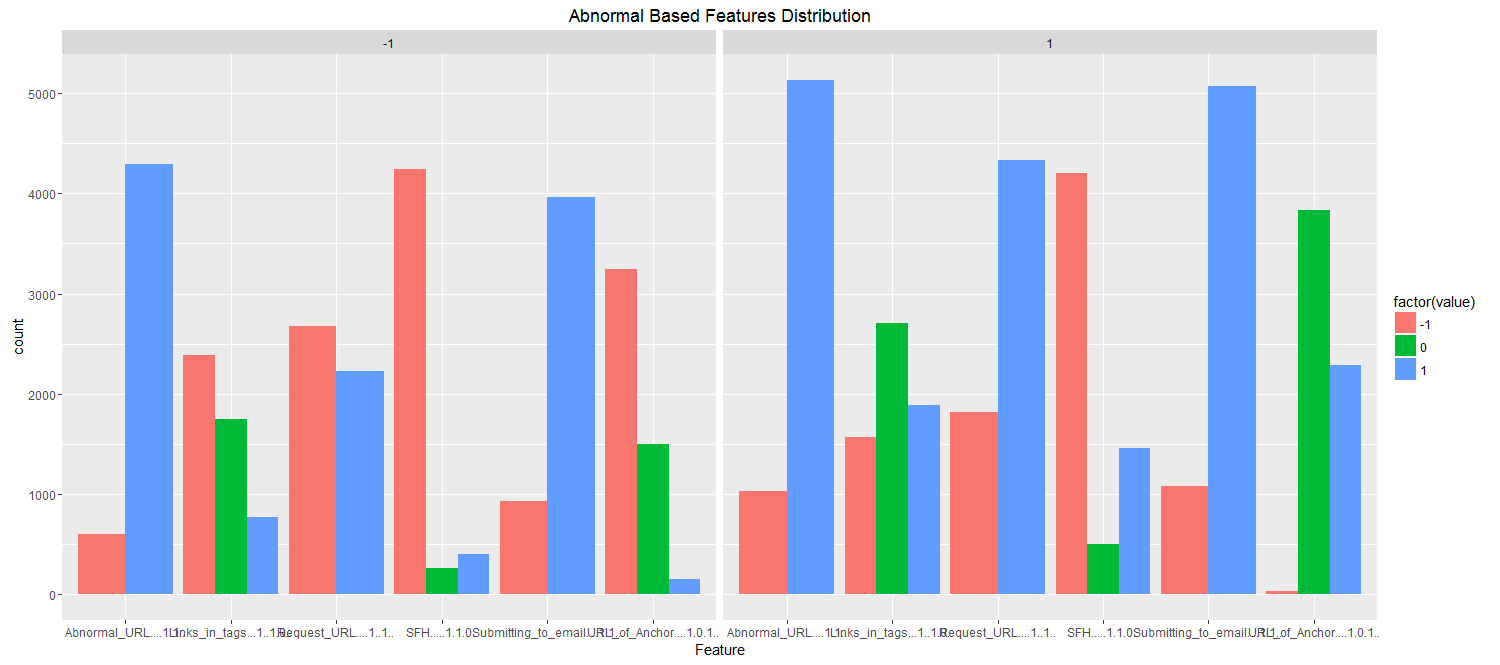
* SSL Final State
* URL of Anchor

**Address Bar Based Features:**

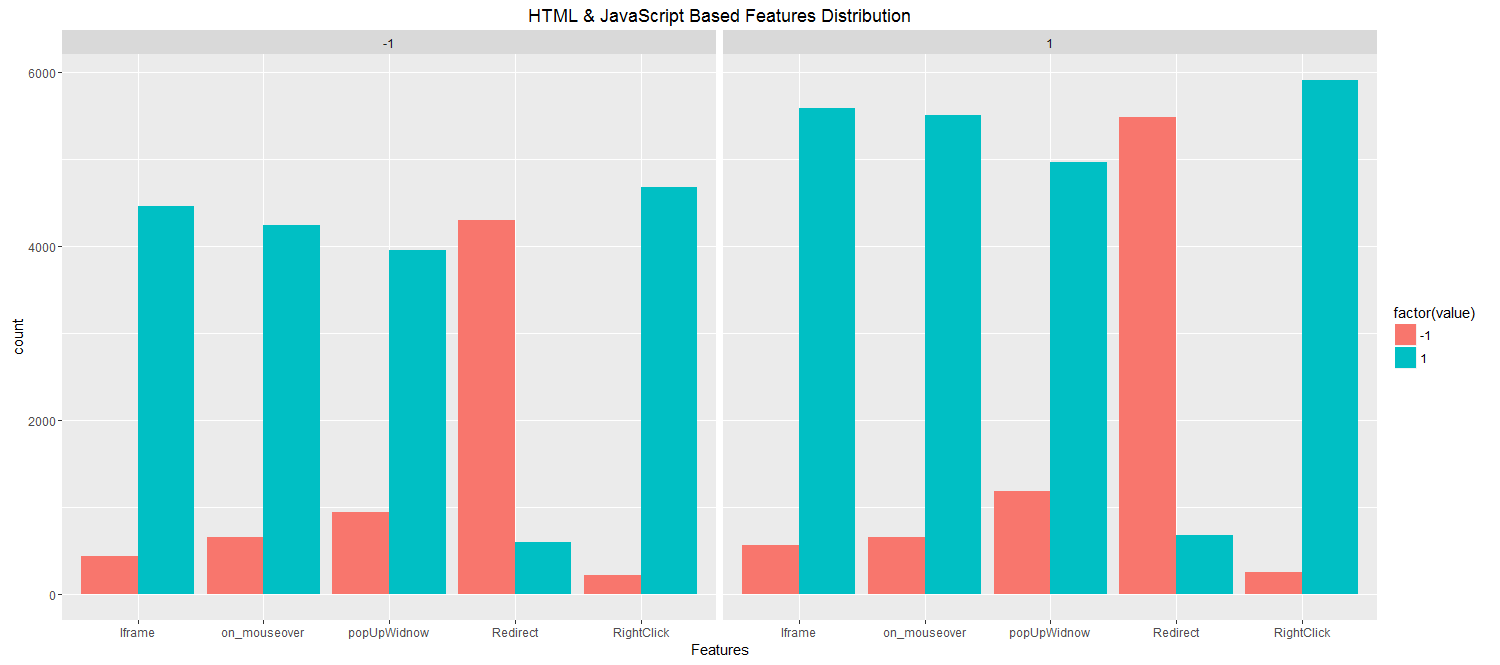
Since there were 12 features I have plotted the features in two separate plots



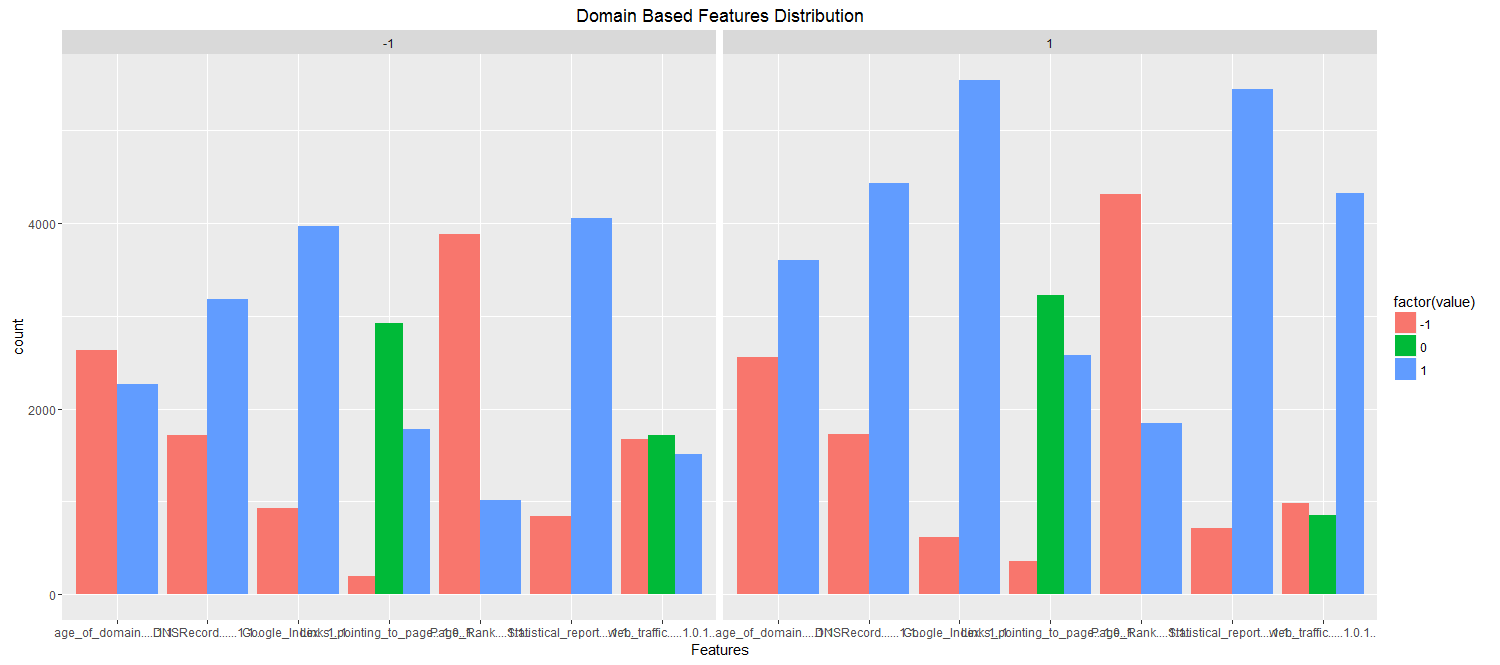


**Abnormal Based Features:**

**HTML and JavaScript based Features:**



**Domain based Features:**

****

#### Model:

In this we did a comparison study of three classification models:

* Logistic Regression
* Random Forest
* Decision Trees

Initially the dataset was divided into equal parts using the SPLIT function. One half is used as training dataset and the other half for testing the model. The same training dataset is used for all the models.

The dependent variable which will be used for the initial model analysis is "Result", this variable which tells us whether a website is legitimate or phishing. "1" means the website is legitimate while "0" means the website is phishing. The independent variables will be added around this dependent variable and re-iterated until a successful model has been created.

Once the models have been built, the "Test" dataset will be used for its evaluation.

#### Results:

## Measuring the accuracy of the model

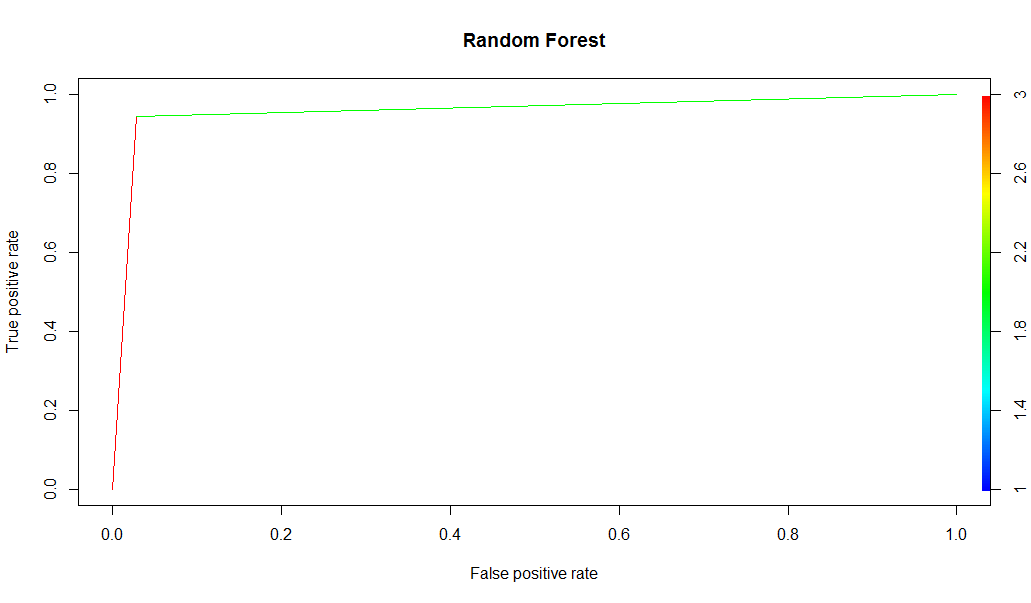
|  |  |  |  |
| --- | --- | --- | --- |
|  | Sensitivity | Specificity | Accuracy |
| Logistic Regression | 94.64% | 90.07% | 92.612% |
| Decision Trees | 91.13% | 89.75% | 90.521% |
| Random Forests | 97.11% | 94.37% | 95.894% |

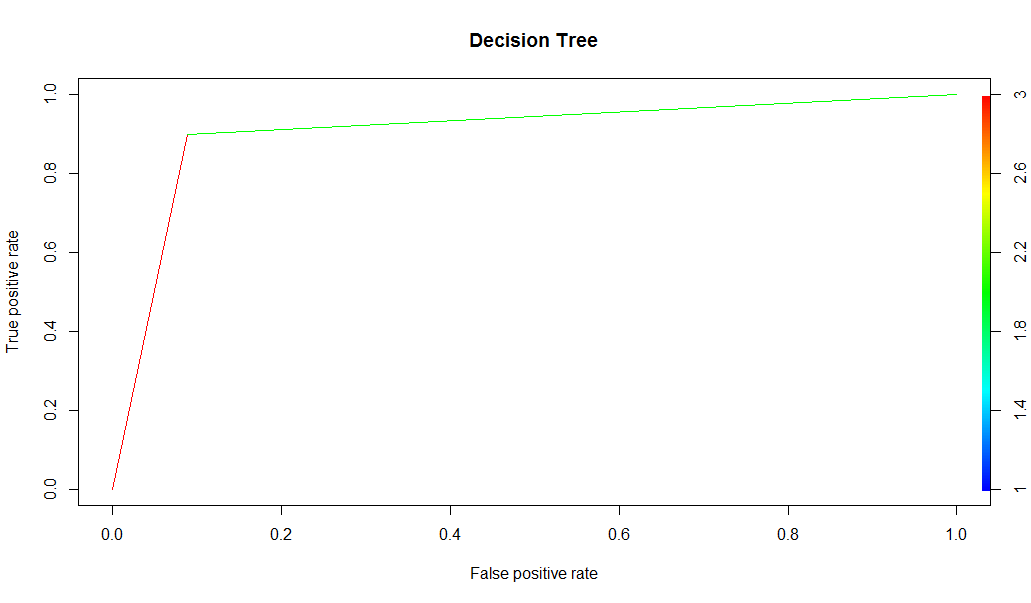
By comparing the accuracy from the above table we can say that Random Forests predicts more accurately compared with the decision trees and logistic regression.

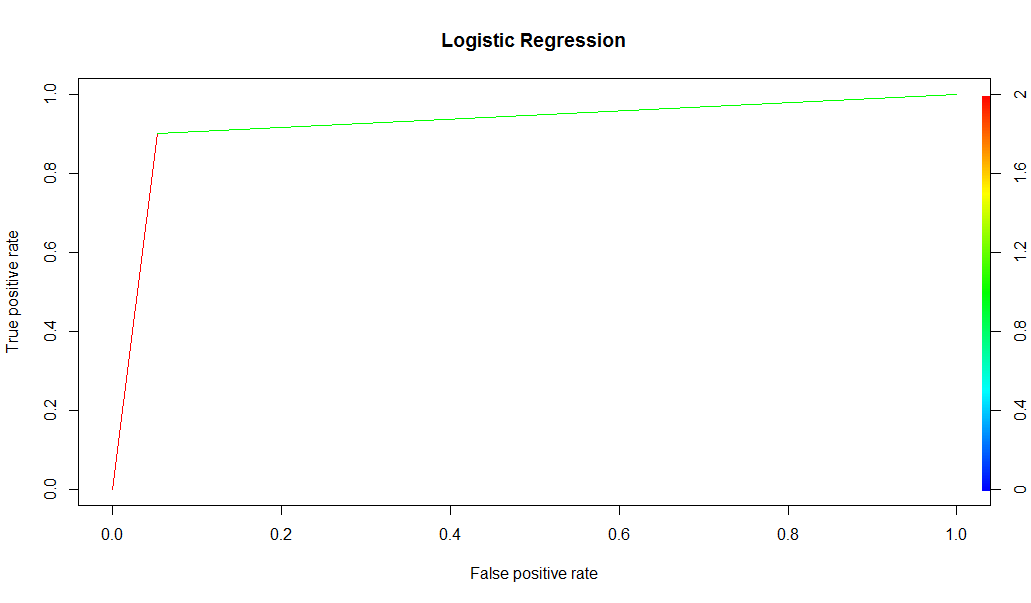
## Threshold Value

In order to investigate the t value further, a good place to start is to determine the area under the ROC curve for the model. The resulting AUC value is an indicator of the discrimination ability of the model. The value ranges from 0.5 to 1 with a higher value representing a better discrimination ability and hence a better predictor.

For example, an ROC curve with an AUC of 0.5 will have a curve close to 45 degrees and will have little or no discrimination ability i.e. the prediction will be 50/50 - pure chance. Whereas a curve that tends towards the top left hand corner of the plot will have an AUC of closer to 1 which indicates perfect discrimination ability. Hence, a high AUC value is more desirable as it indicate a better predictor power for the model in question.





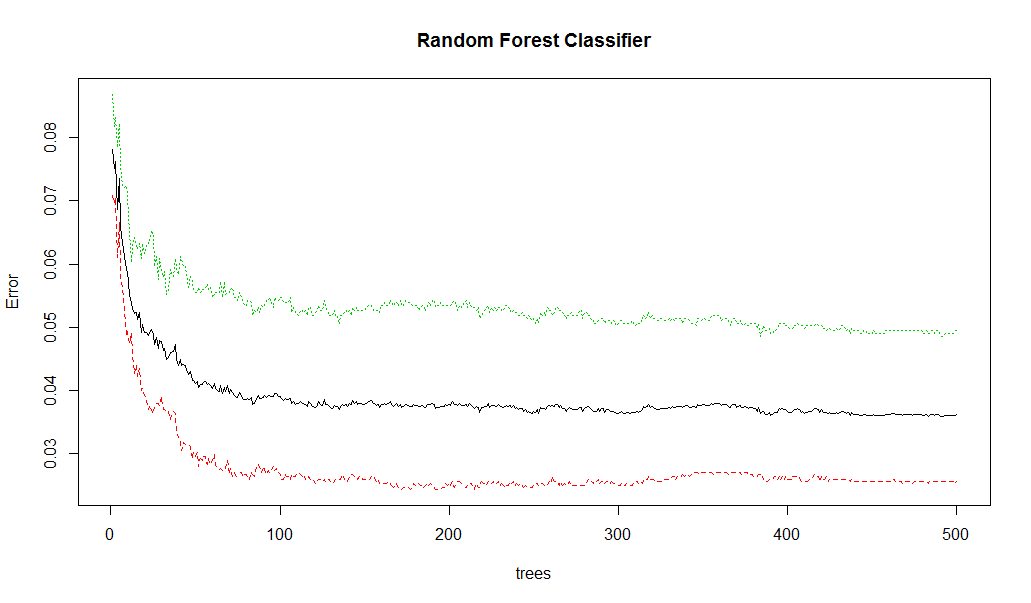


#### Conclusion :

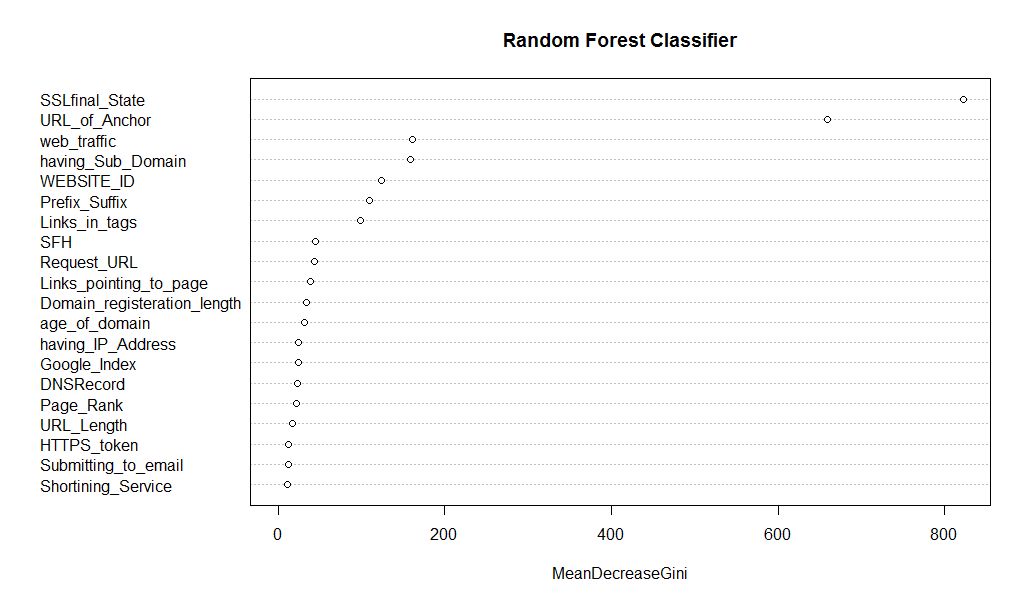
The models which have been built for predicting the phishing websites have been successfully applied to the data. The Random Forest has outperformed the decision tree and logistic regression in terms of accuracy with an optimal sensitivity of 97.11% achieved.

As discussed in above section the t-value, the sensitivity of the model was very important and needed to be as high as possible. It is far more important for the model to keep False Negatives as low as possible rather than False Positives, while still keeping the accuracy high. Hence, there was a trade-off between the 3 main metrics: Accuracy, Sensitivity and Specificity.

We can also optimize the random forest model by setting the no of trees and depth of the trees, which plays a very important role in model accuracy.



By using Random Forests we can also find which features are important and which features can be discarded so that the accuracy of the model can be improved.



From the above plot and from data exploration bar chats we can conclude that SSL final state and URL of Anchor features play very important role in predicting if a website is legitimate or suspicious.