Features Distribution

In this I was trying to see a pattern using individual features, like in case of legitimate websites I was checking which feature had more correlation with the dependent variable. As discussed I have classified the features 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

And throughout this project below were the naming conventions used.

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

Process:

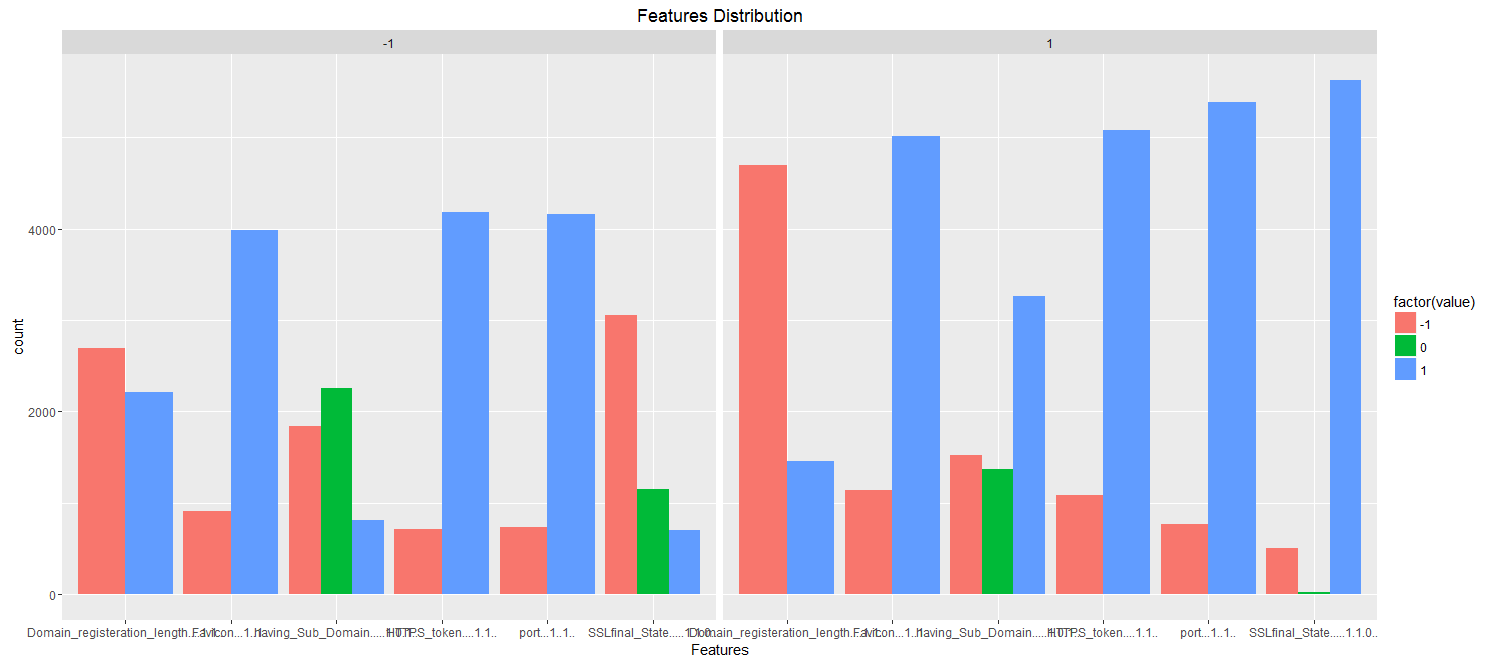
First I have divided the 11000 websites into two categories "Legitimate" and "Phishing" websites depending on the dependent variable. As shown below, Left side bar plots were feature distributions of Phishing websites and the right side bar plots for Legitimate websites.

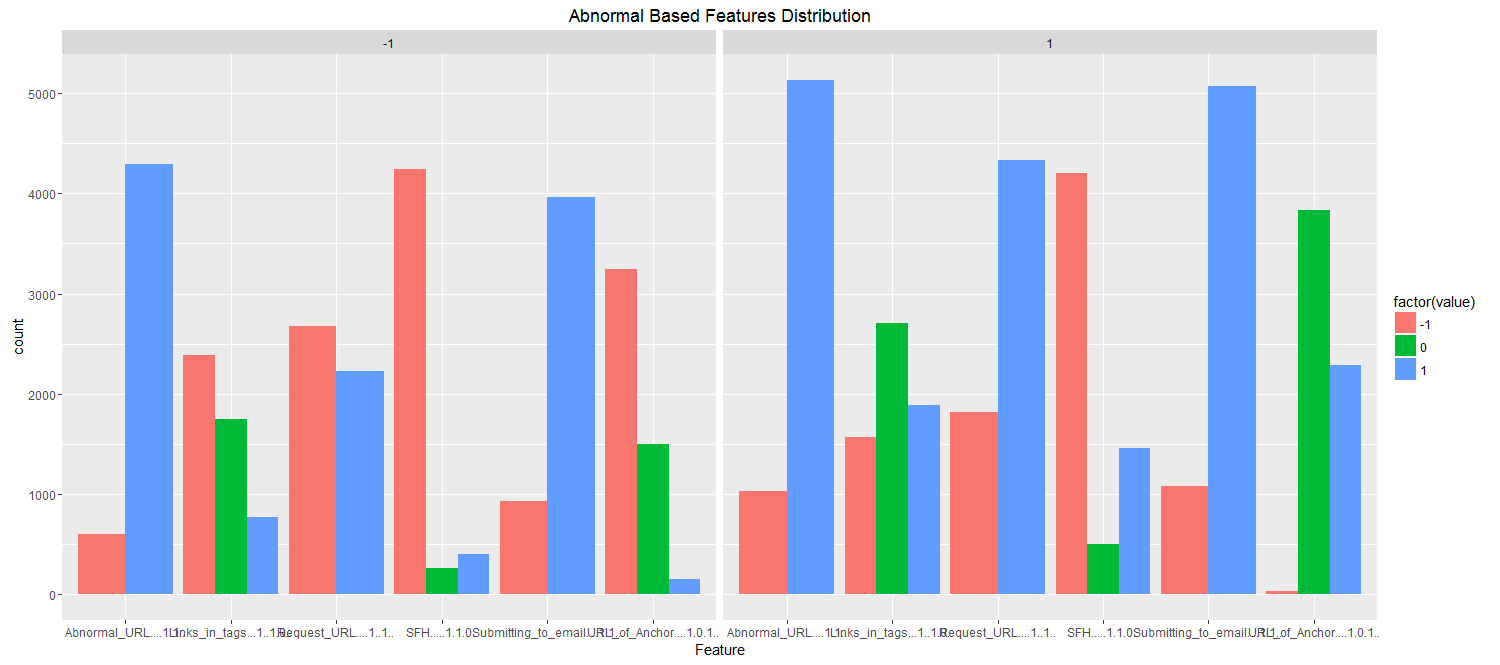
From the below plots I was trying to see which feature was more correlated to the predicting variable.

**Address Bar Based Features:**

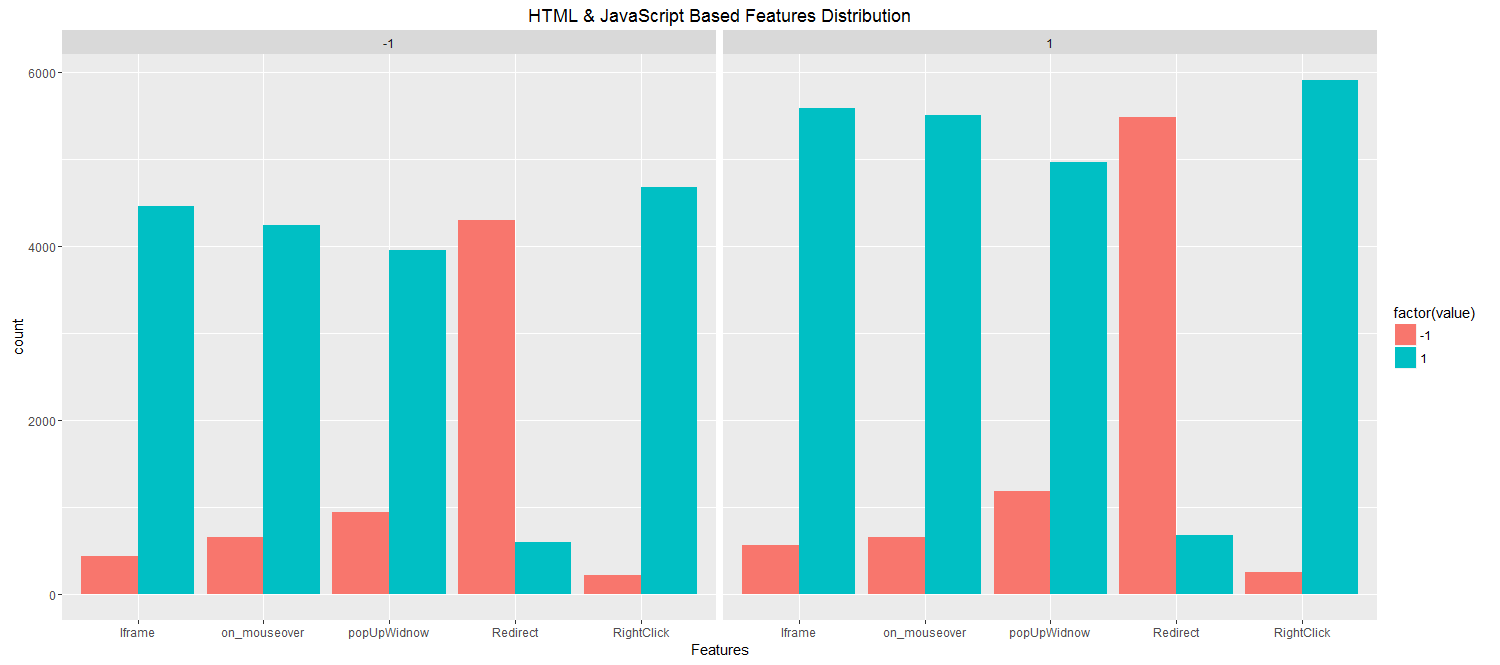
**Since there were 12 features I have plotted the features in two separate plots**



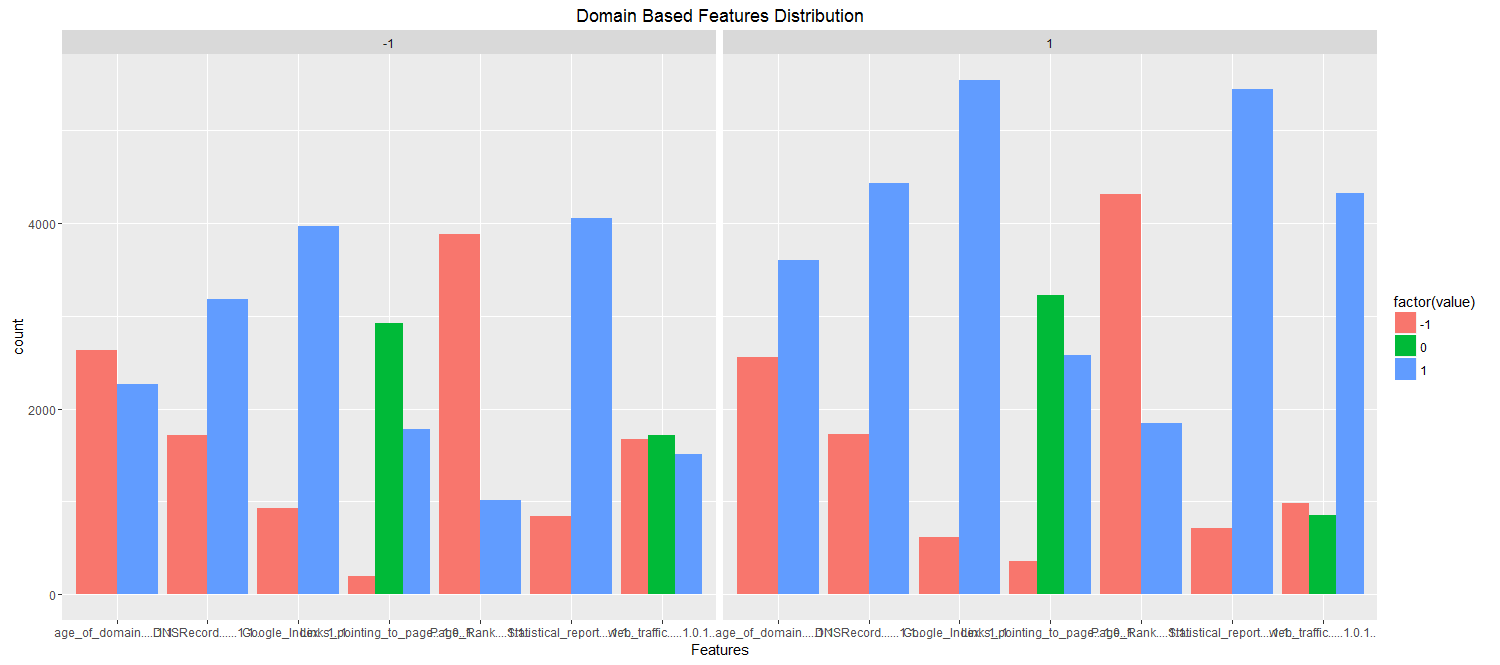


**Abnormal Based Features:**

**HTML and JavaScript based Features:**



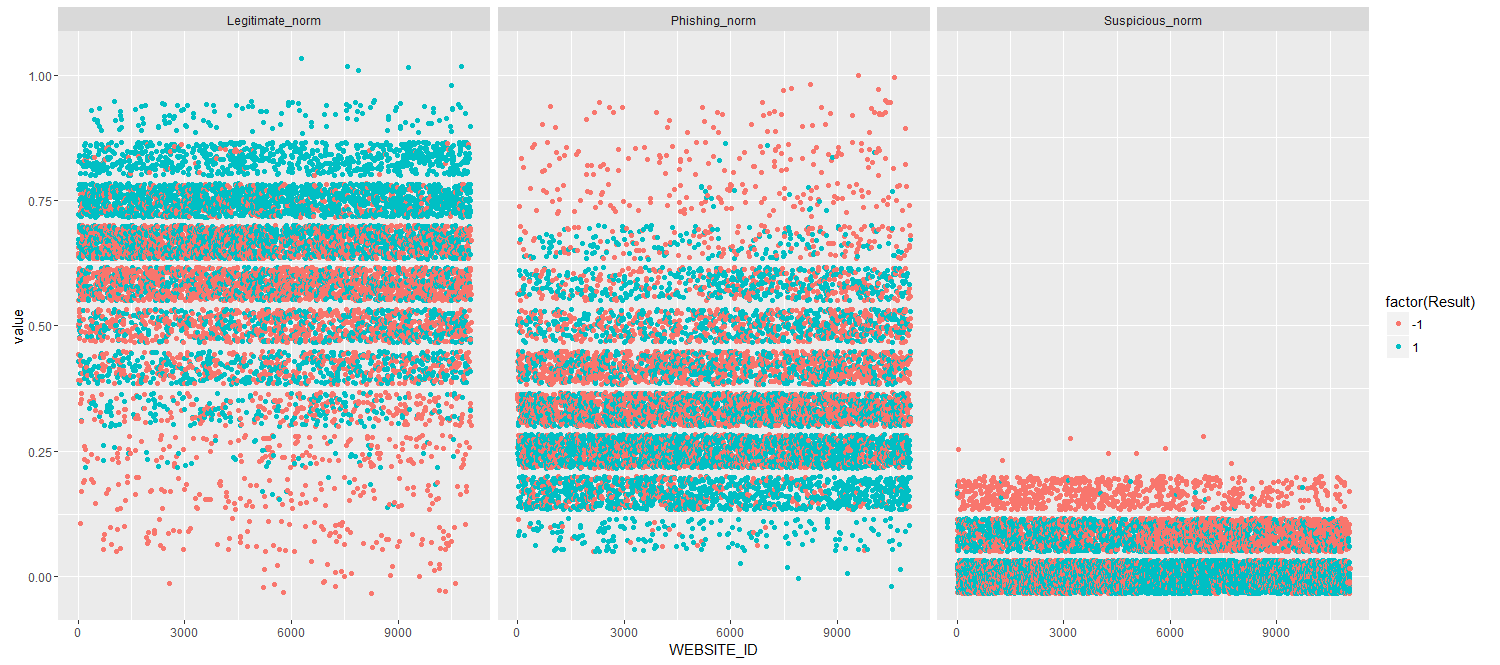
**Domain based Features:**

****

Each Website Features Distributions

Another analyses I did was, for each website there were 30 features and depending on these features the website will be classified as Legitimate or Phishing. So for each website I have calculated the percentage no of the features that consider the website was legitimate , suspicious and phishing and plotted them as a scattered plot. In the below plots Green dots were "Legitimate Websites" and Red dots were "Phishing Websites" . I have plotted this for each category features and also for the overall features.

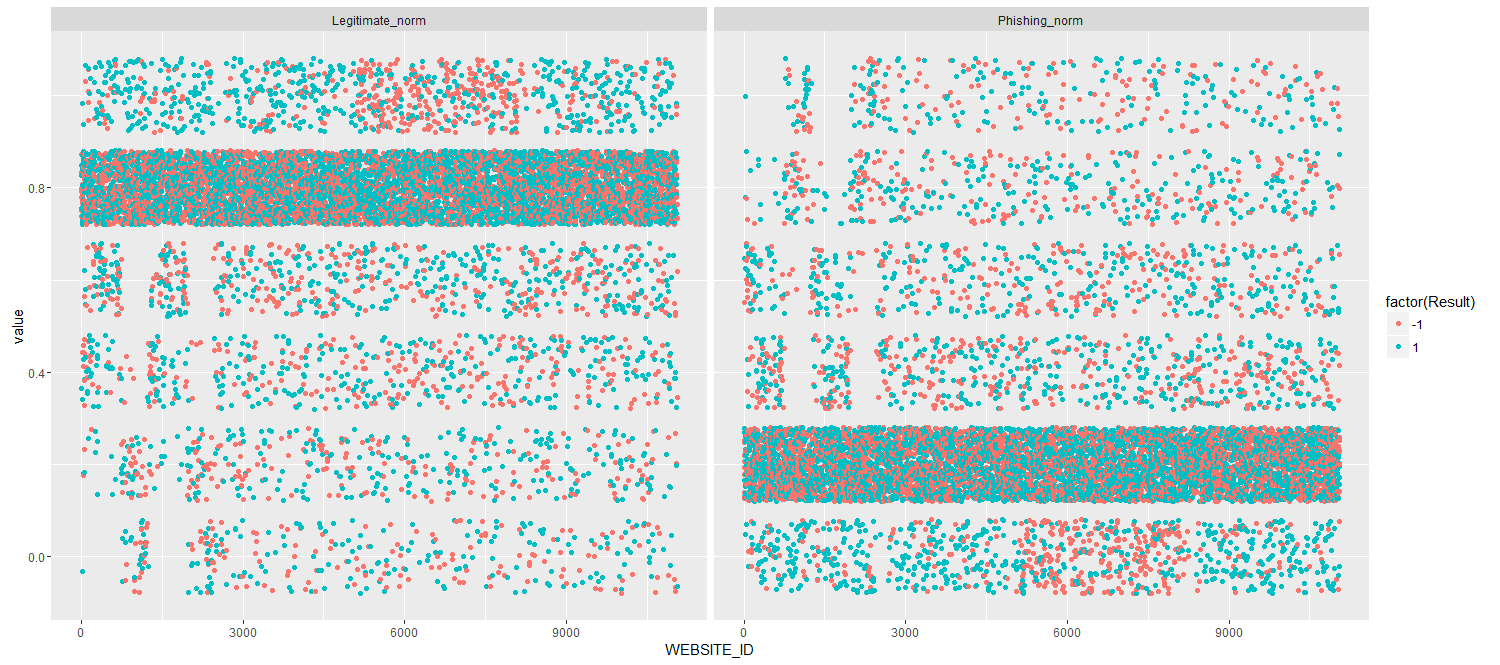
**Address Bar Based Features:**



**Abnormal Based Features:**

****

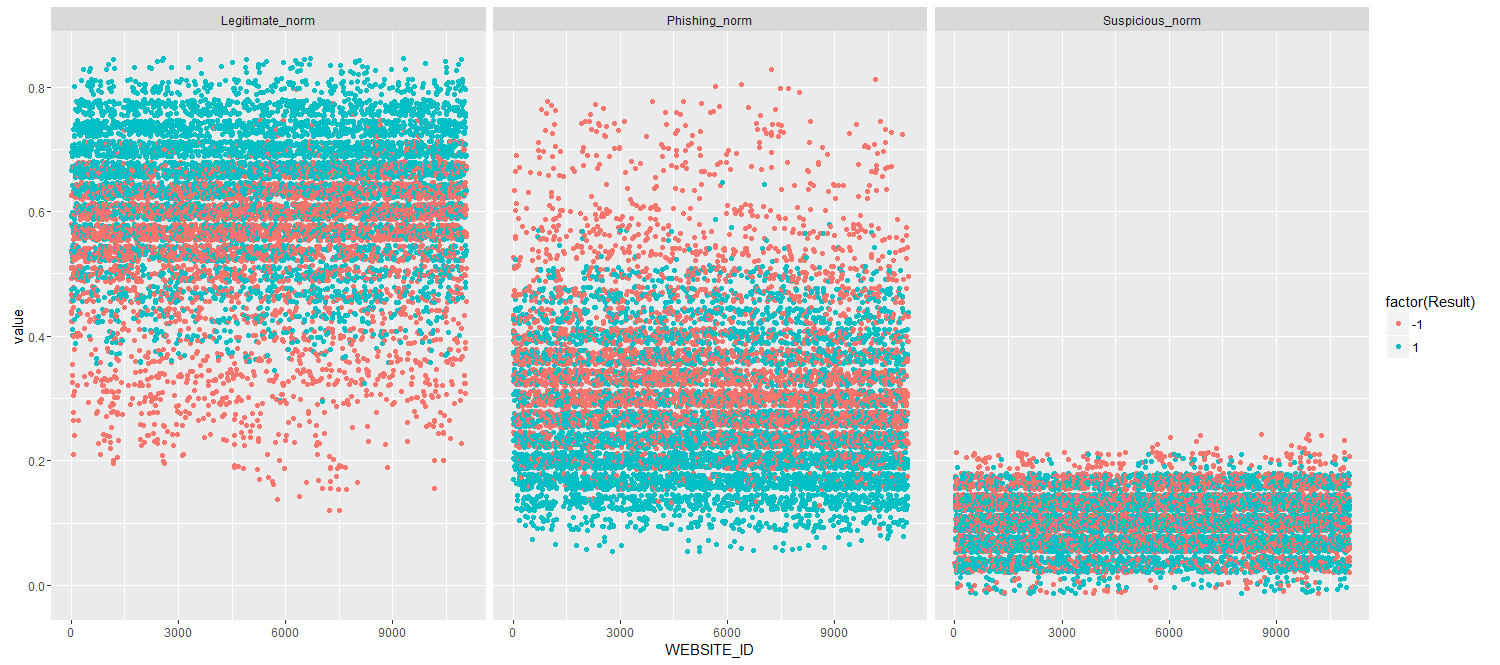
**HTML and JavaScript based Features:**

****

**Domain based Features:**



**Overall Features:**

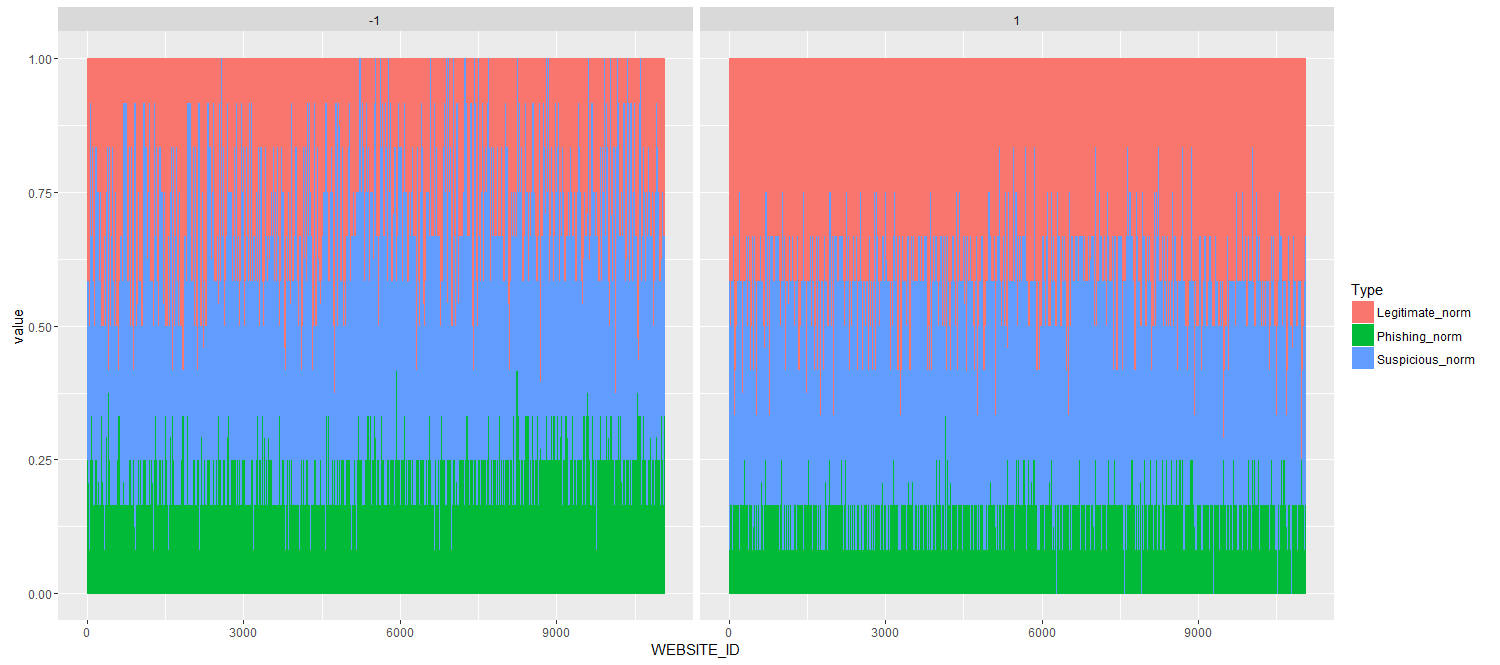


Each Website Features Distributions (Stacked plots)

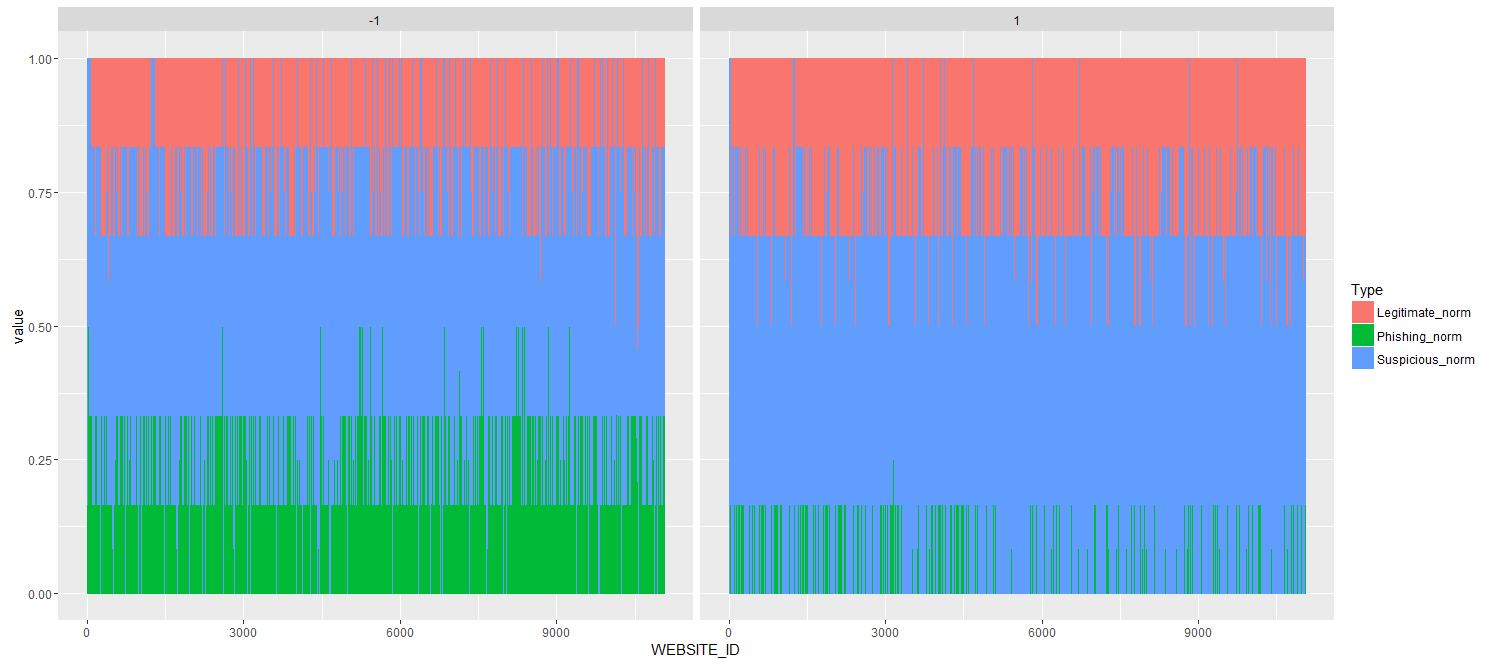
In the above plots for each website we have different plots for features that were predicting legitimate, Suspicious and Phishing.

In this section, I have plotted a stacked view for each website, what percent of the features were predicting legitimate, Suspicious and Phishing. Left side plots were Phishing websites where as the right side plots were Legitimate websites

**Address Bar Based Features:**

****

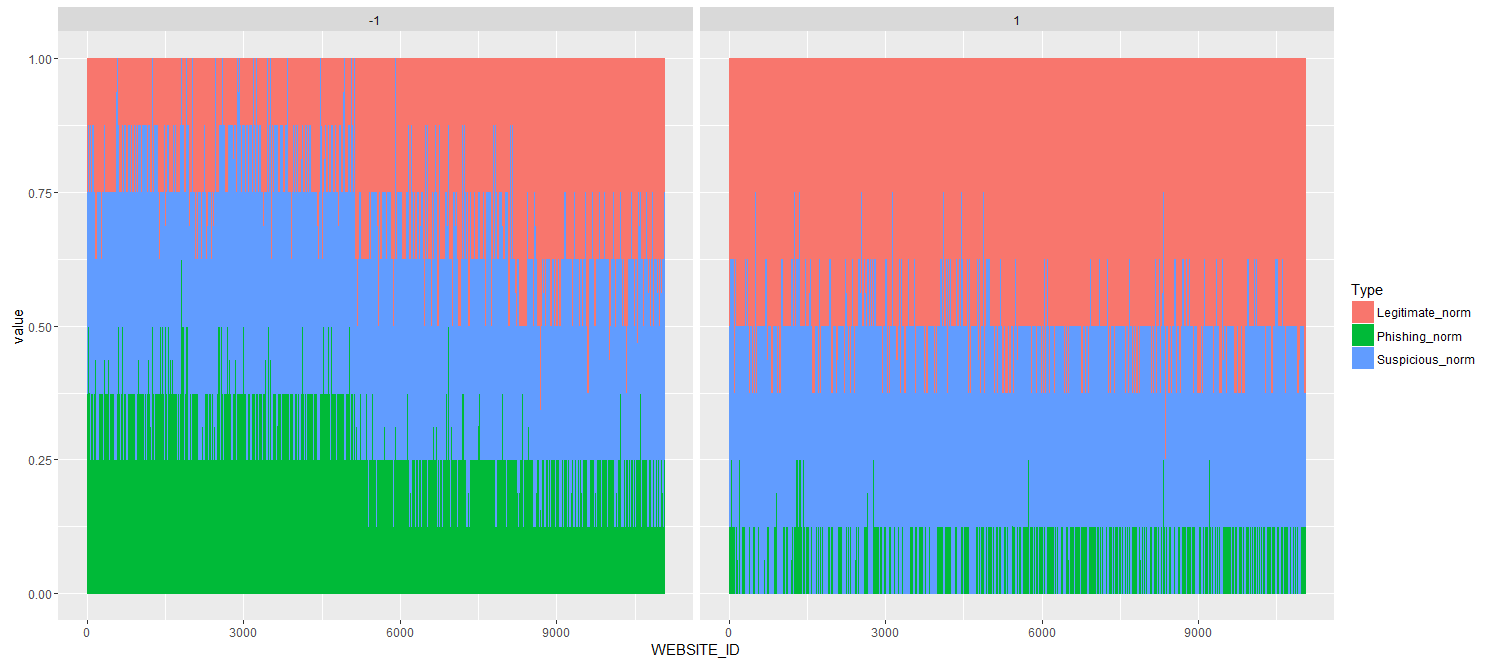
**Abnormal Based Features:**

****

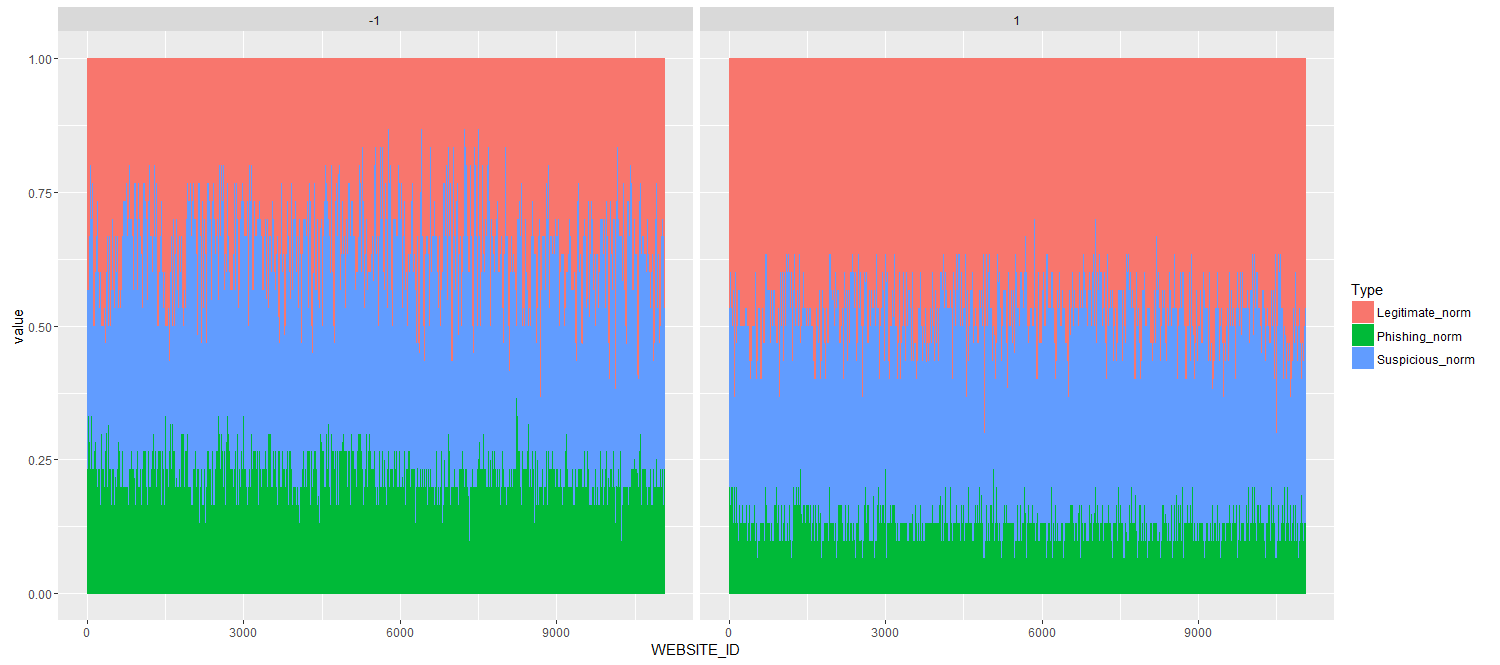
**HTML and JavaScript based Features:**

****

**Domain based Features:**

****

**Overall Features:**

****