Predicting Phishing URLs

CSC-272 Final Project

Jonah Moore, Jason Moore, and Brittany Crawn

1. INTRODUCTION

In today's growing technological world, there is a parallel rise in hacking knowledge. Inboxes today can easily fill up with more spam emails than legitimate emails, and often times, for a common person, it is hard to distinguish the difference between the two, especially with certain programs that create phishing emails and URLs to look almost identical to real ones. Our data set comes from the Donald Bren School of Information and Computer Sciences at the University of California, Irvine. The dataset contains attributes describing characteristics of phishing and non-phishing URLs. The dataset contained 11,000 URLs, none of which had missing values. We split the data set into two files, 70% for training and 30% for testing. The descriptions of the attributes called for a lot of preprocessing for us and other observers to truly understand. We also changed the values of the attributes, from -1, 0 and 1 to phishing, suspicious and legitimate, to be more readable on the surface as well.

With this data set, we aimed to find the most accurate and most telling model to predict whether or not a URL is phishing. In order to do this, we used the data analytics software package Weka (Waikato Environment for Knowledge Analysis) to perform different classification and association algorithms, such as OneR, Naive Bayes, PRISM, IBk, and J48. After running each algorithm to build a model based on the training data, we found that the decision tree was the best indicator for the testing data with the highest overall accuracy and best results of false negatives.

We also created our own small data set consisting of two URLs, mirroring the data set we found, with instances compiled from the recent phishing emails we receive from through myFurman. We tested this data set in the model built from the full training set, hoping to get a relatively high accuracy. After testing each algorithm on this data set, we were able to get 100% on all algorithms.

2. DATA DESCRIPTION

Overall Data Description. The dataset we plan to use consists of the text mining of URLs in emails to determine if they are phishing URLs or not. To further this research, we will use text mining in Weka (or possibly use an outside source) to determine if a URL in an email is phishing or not. The dataset includes attributes such as URL length, having an at symbol, and DNS record that will be used to classify if the URL is phishing or not. The dataset we found includes 11,055 instances, which we will split into two files for training and testing. We will split the data 70/30 for training and testing, respectively.

Attribute Descriptions.

having_IP_Address { phishing, legitimate }

- IP addresses can be utilized as alternatives to domain names in a URL, e.g. http://125.98.3.123/fake.html
- **Rule:** If the Domain Part has an IP Address → Phishing
  Otherwise → Legitimate

URL_Length { legitimate, suspicious, phishing }

- Sometimes, phishers use strangely long URL's to hide the malicious portion of the link, e.g. "http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.website.html"
- Ranges for length were set based on average length of URL's in the original dataset.
- **Rule:** If URL length < 54 → feature = Legitimate
  Else if URL length >= 54 AND <= 75 → feature = Suspicious
  Otherwise → feature = Phishing

Shortining_Service { legitimate,phishing }

- URL shortening has almost no place on the INternet BESIDES attempts to hide malicious sites. Youtube is the only platform that I know of that uses this type of service legitimately
- **Rule:** If TinyURL → Phishing
  Otherwise → Legitimate

having_At_Symbol { legitimate, phishing }

- The '@' character is reserved in most browsers and is only allowed to be used if it is encoded/sanitized. URL's containing @ symbols are rarely legitimate.
- **Rule:** If URL having @ symbol → Phishing

Otherwise → Legitimate

double_slash_redirecting { phishing,legitimate }

- The existence of "//" within the URL path means that the user will be redirected to another website. An example is "http://www.legitimate.com//http://www.phishing.com". Redirects are rarely necessary in a link and multiple redirects usually suggest a link is malicious.
- **Rule:** If the position of the last occurrence of "//" in the URL > 7 → Phishing
  Otherwise → Legitimate

Prefix_Suffix { phishing, legitimate }

- Dashes are not a reserved symbol, but they are generally left out of URLs and are considered "bad practice" to include. Phishers utilize dashes to include familiar names in their URL, i.e. "http://www.confirme-paypal.com/"
- **Rule:** If Domain name part includes (-) symbol → Phishing
  Otherwise → legitimate

having_Sub_Domain { phishing,suspicious,legitimate }

- This section of the dataset was heavily parsed in order to determine the number of "irrelevant" subdomains in a URL. The creators of the dataset removed the "www" subdomain and removed all top level domains (edu, com, net, etc.). The remaining dots were counted and used to form the rule below.
- **Rule:** If Dots in domain part = 1 → legitimate
  Else if dots in domain part = 2 → suspicious
  Otherwise → phishing

SSLfinal_State { phishing,legitimate,suspicious }

- The use of HTTPS is very important in determining a website's legitimacy. Certificates used to generate the Secure Socket Layer (SSL) encryption used by HTTPS are issued by centralized, trusted certificate authorities.
- **Rule:** If Use HTTPS and Issuer is Trusted and Age of Certificate >= 1 years →
  legitimate
  Else if using HTTPS and Issuer is Not Trusted → suspicious
  Otherwise → phishing

Domain_registeration_length { phishing,legitimate }

- In many cases, reliable domain names are registered years in advance or have been registered for a long period of time. Therefore, brand new domain names are generally questionable.
- **Rule:** If Domains expires on <= 1 years → phishing
  Otherwise → legitimate

Favicon { legitimate,phishing }

- The small image in the tab at the top of your browser is the "favicon." Reliable/legitimate websites store the favicon on their server and send it with the initial "GET" request for the page. It is very suspicious if the domain loads a favicon from an external source.
- **Rule:** If Favicon loaded from external domain → phishing
  Otherwise → legitimate

port { legitimate,phishing }

- Some ports are uncommon to use for simple web server tasks. Below is a table describing some of the common/desirable port configurations. If a domain is trying to use a port other than the HTTP or HTTPS standard, it is sometimes considered suspicious, ESPECIALLY in the context of an email link.

PORT	Service	Meaning	Preferred Status
21	FTP	Transfer files from one host to another	Close
22	SSH	Secure File Transfer Protocol	Close
23	Telnet	provide a bidirectional interactive text-oriented communication	Close
80	HTTP	Hyper test transfer protocol	Open
443	HTTPS	Hypertext transfer protocol secured	Open
445	SMB	Providing shared access to files, printers, serial ports	Close
1433	MSSQL	Store and retrieve data as requested by other software applications	Close
1521	ORACLE	Access oracle database from web.	Close
3306	MySQL	Access MySQL database from web.	Close
3389	Remote Desktop	allow remote access and remote collaboration	Close

- **Rule:** If port # is of the preferred status → phishing
  Otherwise → Legitimate

HTTPS_token { phishing,legitimate }

- An https token can be included in the domain portion of a URL as an attempt to trick the user into believing it is reliable.
- **Rule:** If using HTTP token in domain part of the URL → phishing
  Otherwise → legitimate

Request_URL { legitimate,phishing }

- Often times a web page will load external resources, such as images, videos and sounds, but it will share the same domain. Malicious sites tend to load these resources from external sources that reside on different domains.
- **Rule:** If % of request URL < 22% → legitimate
  Else if % of request URL >= 22% AND <= 61% → suspicious
  Otherwise → feature = phishing

URL_of_Anchor { phishing,suspicious,legitimate }

- Anchor refers to the <a> tag in this context. If an anchor takes up a larger portion of the URL, it is generally considered more suspicious.
- **Rule:** If % of URL of Anchor < 31% → legitimate
  Else if % of URL of anchor >= 31% and<= 67% → Suspicious
  Otherwise → phishing

Links_in_tags { legitimate,phishing,suspicious }

- If large portions of a link are enclosed in tags, it is considered suspicious.
- **Rule:** If % of links in "<Meta>," "<Script>", and "<Link>" < 17% → legitimate
  Else if % of links in "<Meta>," "<Script>", and "<Link>" >=17% and <=81% → Suspicious
  Otherwise → Phishing

SFH { phishing,legitimate,suspicious }

- If a form is being submitted to "about:blank" or to nowhere (empty string), then the link is likely malicious. It is also suspicious if the SFH points to an external domain because most forms are not sent to different domains.
- **Rule:** If SFH is "about: blank" Or is Empty → phishing
  SFH Refers to a different domain → suspicious
  Otherwise → legitimate

Submitting_to_email { phishing,legitimate }

- If a form submission is directed to an email account, it is likely that the link is malicious.
- **Rule:** If Using "mail()" or "mailto:" Function to Submit User information → phishing
  Otherwise → legitimate

Abnormal_URL { phishing,legitimate }

- Apparently, there is a WHOIS database that contains information about if a host name is included in the URL. It was referenced in the creation of this dataset.
- **Rule:** If the hostname is not included in URL → Phishing
  Otherwise → Legitimate

Redirect { suspicious,legitimate }

- Legitimate sites are rarely redirected more than once. Plain and simple.
- **Rule:** If ofRedirect Page <=1 → legitimate
  Else if ofRedirect Page >=2 AND < 4 → suspicious
  Otherwise → Phishing

on_mouseover { legitimate,phishing }

- The "onMouseOver" function has to ability to change the information displayed in the status bar when the link is hovered over. If the status bar content is changed by this method, it is usually for malicious purposes.
- **Rule:** If onMouseOver changes status bar → Phishing

Otherwise → legitimate

RightClick { legitimate,phishing }

- Some phishing sites will disable right click capability so the source code of the site is not viewable. If the string "event.button==2" appears anywhere in the source code, it is likely trying to disable right click.
- **Rule:** IF Right click disabled → phishing
  Otherwise → legitimate

popUpWindow { legitimate,phishing }

- Most pop-up and alert boxes are not used to create forms or take user input. If alert boxes are used for this purpose, it is a probably a malicious attempt to steal information.
- **Rule:** If Popup window contains text fields → Phishing
  Otherwise → legitimate

Iframe { legitimate,phishing }

- **Rule:** If using iframe → Phishing
  Otherwise → legitimate

age_of_domain { phishing,legitimate }

- This can also be extracted from the WHOIS database that was previously mentioned. If a site lives for fewer than 6 months, it is likely a phishing site.
- **Rule:** If Age of domain >= 6 months → legitimate
  Otherwise → phishing

DNSRecord { phishing,legitimate }

- This feature is also extracted from the WHOIS database. Self explanatory rules.
- **Rule:** If no DNS Record for the domain → phishing
  Otherwise → legitimate

web_traffic { phishing,suspicious,legitimate }

- This feature seems questionable to me and it might be worthwhile to remove it from the dataset.
- Rule: If website rank <100,000 → Legitimate
  Else if website rank > 100,000 → Suspicious
  Otherwise → phishing

Page_Rank { phishing,legitimate }

- PageRank scores range from a value of "0" to "1". PageRank uses a sites "connectedness," or the number of links routing to and from the site, to determine its relevance to Google searches.
- **Rule:** If PageRank < 0.2 → Phishing
  Otherwise → legitimate

Google_Index { legitimate,phishing }

- This rule is fairly self explanatory.
- **Rule:** If Webpage indexed by Google → legitimate
  Otherwise → phishing

Links_pointing_to_page { legitimate,suspicious,phishing }

- I don't know how they were able to measure this. It seems questionable as well and may be worth removing from the dataset, as this is already covered relatively well by the PageRank score.
- **Rule:** If Of Linking pointing to the Webpage = 0 → Phishing
  Else if Of Linking to the Webpage > 0 and <= 2 → Suspicious
  Otherwise → legitimate

Statistical_report { phishing,legitimate }

- There are a variety of third party groups that index known phishing websites and make them publicly available. If the site/link is indexed by one of these companies, then it is KNOWN to be malicious.
- **Rule:** If host belongs to Top Phishing IPs or Top Phishing Domains → Phishing
  Otherwise → legitimate

Result { phishing,legitimate }

- This is the class attribute of our dataset.
- **Rule:** If URL was phishing → phishing
  Otherwise → legitimate

3. DATA PREPARATION

Data Set Prep. The current dataset we found used -1, 0, and 1 for the values of each attribute to describe phishing, suspicious, and legitimate, respectively [Fig. 1]. This is not as descriptive as we would like the raw data to be for our visualizations. However, when looking at the data and the associated descriptions given in a different document, the descriptions did not tell us what each number described. We had to analyze the data and the descriptions to determine the correct relationship between the number and the actual descriptive value of phishing, suspicious and legitimate. So, once we determined that -1 is phishing, 0 is suspicious, and 1 is legitimate for all attributes (even ones with just 0 and 1 or -1 and 1), we used *Find All* and *Replace* to change all the numbers to their respective nominal description, as seen in Fig. 2. Lastly, we fixed some misspellings in the attribute names.

```
@attribute popUpWidnow  { 1,-1 }
@attribute Iframe { 1,-1 }
@attribute age_of_domain  { -1,1 }
@attribute DNSRecord   { -1,1 }
@attribute web_traffic  { -1,0,1 }
@attribute Page_Rank { -1,1 }
@attribute Google_Index { 1,-1 }
@attribute Links_pointing_to_page { 1,0,-1 }
@attribute Statistical_report { -1,1 }
@attribute Result  { -1,1 }


@data
-1,1,1,1,-1,-1,-1,-1,-1,1,1,-1,1,-1,1,-1,-1,-1,0,1,1,1,1,-1,-1,-1,-1,1,1,-1,-1
1,1,1,1,1,-1,0,1,-1,1,1,-1,1,0,-1,-1,1,1,0,1,1,1,1,-1,-1,0,-1,1,1,1,-1
1,0,1,1,1,-1,-1,-1,-1,1,1,-1,1,0,-1,-1,-1,-1,0,1,1,1,1,1,-1,1,-1,1,0,-1,-1
1,0,1,1,1,-1,-1,-1,1,1,1,-1,-1,0,0,-1,1,1,0,1,1,1,1,-1,-1,1,-1,1,-1,1,-1
1,0,-1,1,1,-1,1,1,-1,1,1,1,1,0,0,-1,1,1,0,-1,1,-1,1,-1,-1,-1,0,-1,1,1,1,1
-1,0,-1,1,1,-1,-1,1,1,-1,1,1,-1,1,0,0,-1,-1,-1,0,1,1,1,1,1,1,1,-1,1,-1,-1,1
1,0,-1,1,1,-1,-1,-1,1,1,1,1,-1,-1,0,-1,-1,-1,0,1,1,1,1,1,1,-1,-1,-1,1,0,-1,-1
1,0,1,1,1,-1,-1,-1,1,1,1,-1,-1,0,-1,-1,1,1,0,1,1,1,1,-1,-1,0,-1,1,0,1,-1
1,0,-1,1,1,-1,1,1,-1,1,1,-1,1,0,1,-1,1,1,0,1,1,1,1,1,-1,1,1,1,0,1,1
1,1,-1,1,1,-1,-1,1,-1,1,1,1,1,0,1,-1,1,1,0,1,1,1,1,1,-1,0,-1,1,0,1,-1
1,1,1,1,1,-1,0,1,1,1,1,1,-1,0,0,-1,-1,-1,0,1,1,1,1,-1,1,1,1,1,-1,-1,1
1,1,-1,1,1,-1,1,-1,-1,1,1,1,1,-1,-1,-1,-1,-1,-1,0,1,1,1,1,-1,-1,-1,-1,1,0,-1,-1
```

Figure 1. A sample of the original data set with -1, 0, 1 as the attribute values. You can also see a misspelling of “popUpWindow” in the attribute names.

```
@attribute popUpWindow  { legitimate,phishing }
@attribute Iframe { legitimate,phishing }
@attribute age_of_domain  { phishing,legitimate }
@attribute DNSRecord   { phishing,legitimate }
@attribute web_traffic  { phishing,suspicious,legitimate }
@attribute Page_Rank { phishing,legitimate }
@attribute Google_Index { legitimate,phishing }
@attribute Links_pointing_to_page { legitimate,suspicious,phishing }
@attribute Statistical_report { phishing,legitimate }
@attribute Result  { phishing,legitimate }


@data
phishing,legitimate,legitimate,legitimate,phishing,phishing,phishing,phishing,phishing,legitimate,legitimate,phishing,legitimate,phishing,legitimate,phishing,ph
    ishing,phishing,suspicious,legitimate,legitimate,legitimate,legitimate,phishing,phishing,phishing,phishing,legitimate,legitimate,phishing,phishing
legitimate,legitimate,legitimate,legitimate,legitimate,phishing,suspicious,legitimate,phishing,legitimate,legitimate,phishing,legitimate,suspicious,phishing,phi
    shing,legitimate,legitimate,suspicious,legitimate,legitimate,legitimate,legitimate,phishing,phishing,suspicious,phishing,legitimate,legitimate,legitimate,ph
    ishing
legitimate,suspicious,legitimate,legitimate,legitimate,phishing,phishing,phishing,phishing,legitimate,legitimate,phishing,legitimate,suspicious,phishing,phishin
    g,phishing,phishing,suspicious,legitimate,legitimate,legitimate,legitimate,legitimate,phishing,legitimate,phishing,legitimate,suspicious,phishing,phishing
legitimate,suspicious,legitimate,legitimate,legitimate,phishing,phishing,phishing,legitimate,legitimate,legitimate,phishing,phishing,suspicious,suspicious,phish
    ing,legitimate,legitimate,suspicious,legitimate,legitimate,legitimate,legitimate,phishing,phishing,legitimate,phishing,legitimate,phishing,legitimate,phishi
    ng
legitimate,suspicious,phishing,legitimate,legitimate,phishing,legitimate,legitimate,phishing,legitimate,legitimate,legitimate,legitimate,suspicious,suspicious,p
    hishing,legitimate,legitimate,suspicious,phishing,legitimate,phishing,legitimate,phishing,phishing,suspicious,phishing,legitimate,legitimate,legitimate,legi
    timate
phishing,suspicious,phishing,legitimate,phishing,phishing,legitimate,legitimate,phishing,legitimate,legitimate,phishing,legitimate,suspicious,suspicious,phishin
    g,phishing,phishing,suspicious,legitimate,legitimate,legitimate,legitimate,legitimate,legitimate,legitimate,phishing,legitimate,phishing,phishing,legitimate
legitimate,suspicious,phishing,legitimate,legitimate,phishing,phishing,phishing,legitimate,legitimate,legitimate,legitimate,phishing,phishing,suspicious,phishin
    g,phishing,phishing,suspicious,legitimate,legitimate,legitimate,legitimate,legitimate,phishing,phishing,phishing,legitimate,suspicious,phishing,phishing
legitimate,suspicious,legitimate,legitimate,legitimate,phishing,phishing,phishing,legitimate,legitimate,legitimate,phishing,phishing,suspicious,phishing,phishin
    g,legitimate,legitimate,suspicious,legitimate,legitimate,legitimate,legitimate,phishing,phishing,suspicious,phishing,legitimate,suspicious,legitimate,phishi
    ng
```

Figure 2. A sample of the revised data set with phishing, suspicious, and legitimate as the attribute values. We also fixed spelling errors.

Data Description Prep. It was obvious that many of the data descriptions were lengthy and confusing. We went through and revised many of the descriptions. We also added the attribute names to each associated description because the original descriptions did not include them. The lack of the attribute names was confusing to us as data analyzers, and it would also be to other observers. Overall, we revised much of the description file in order to make it more understandable to the common reader and to allow us to have an easy document to follow and reference during our data analysis. You can see these descriptions in the Section 2.

4. DATA ANALYSIS

Note: We utilized multiple different forms of validation to test the accuracy of all of our models.

- Tested model using the full training data set (70% of entire data set)
- Tested model using the full supplied test data set (30% of entire data set)
- Tested model using 10-fold cross validation on full training data set
- Tested model using a 66% split on the training data set. This form of testing was more a way to confirm that the splitting of the full data set was random and fairly distributed.

Algorithm #1: OneR

OneR, which stands for "one rule," is a ***classification*** algorithm that generates one rule for each attribute in the data set and simply selects the rule with the highest accuracy (or lowest error rate) for predicting the class attribute. In other words, the algorithm selects the attribute in the data set that is most predictive of the output attribute.

OneR was the first algorithm that we ran on our data set in order to gain a better understanding of which single attribute was most significant for predicting whether or not a URL was phishing or legitimate.

Algorithm #2: Naive Bayes

Naive Bayes is a ***classification*** algorithm based on the Bayesian theorem that utilizes the conditional probabilities derived from the training data set to calculate the likelihood of an outcome being a particular class. The algorithm considers all attributes both independent and of equal importance. It also handles missing values extremely well through use of a Laplace estimator. However, it should be noted that Naive Bayes is not good at handling redundant attributes.

Algorithm #3: PRISM

PRISM is a ***covering*** and ***classification*** algorithm that generates useful rules from the data set until the number of positive examples of a class covered by a rule is equal to the total number of instances covered by that rule (p/t = 1). It is known as a "separate-and-conquer" algorithm because it separates out all instances that a specific rule covers and then "conquers" the remaining instances not already covered. The order of rules generated does NOT matter because all rules are predicting the same class.

Algorithm #4: IBk

The nearest neighbor algorithm is a lazy ***classification*** algorithm that assumes all attributes are equally significant. The algorithm completes a linear scan of all instances in the data set to find the "nearest neighbor" to the instance it is trying to classify. K represents the number of neighbors that are matched with the instance trying to be classified, and a variance in K's size can often change the classification output of the algorithm. We decided to run IBk with K = 1, 3, and 5 to cover multiple possibilities for classification and look at the pattern to determine if we needed to increase K any more..

Algorithm #5: J48

J48 is a decision tree algorithm. In order to create the tree, J48, like many other tree algorithm utilizes the concept of information entropy. Using the training data set, it splits sets of samples into one class or another measuring the information gain and information loss (entropy). Ultimately, the attribute with the highest information gain is the next decision in the tree. We used J48 for our data analysis, changing the confidence level to determine the best overall accuracy.

5. RESULTS

Overall Algorithm Accuracies

Algorithm	Full Training Data Set (70%) Tested on Itself	10-fold Cross Validation on Training Set	66% Split on Training Set	Test Data* (30%)
OneR	88.8745%	88.8745%	89.1676%	**88.9324%**
Naive Bayes	92.919%	92.8027%	93.1585%	**91.0736%**
PRISM	98.21%	96.8%	96.23%	**88.23%**
IBk (k = 1)	99.005%	97.4932%	97.2245%	**91.3752%**
IBk (k = 3)	97.7%	96.0589%	95.401%	**90.4403%**
IBk (k = 5)	96.7955%	95.2319%	94.9829%	**90.1689%**
J48, (confidence = 0.75)	98.5915%	96.7567%	96.8073%	**91.2847%**

Table 1. Accuracies of algorithms tested using both our training and test data sets.

Analysis of Each Algorithm Used

OneR

The attribute selected by the OneR algorithm as most predictive of whether a URL was phishing or legitimate was SSLfinal_State.

SSLfinal_State:

Phishing → Result = phishing
Legitimate → Result = legitimate
suspicious→ Result = phishing

As stated earlier, certificates used to generate the Secure Socket Layer (SSL) encryption used by HTTPS are issued by centralized, trusted certificate authorities. Therefore, we were not surprised that this attribute was most significant in predicting whether a URL was phishing or legitimate given the trust behind the certificate issuing authorities. We used this overall accuracy as a baseline. The accuracy of correctly predicting a phishing URL with this algorithm is **85.7%**.

	Predicted phishing	**Predicted legitimate**
Actual Phishing	1255	208
Actual Legitimate	159	1694

Confusion Matrix (trained using full training set (70%), tested using full test set (30%))

OneR: Removing Attributes

Out of curiosity, we removed the *SSLfinal_State* attribute in both the training data set (70%) and testing data set (30%) to see what the next best predictor of phishing URLs was. When we did this, URL_of_anchor was selected by OneR as the most significant attribute (accuracy = 85%). We repeated the removal process again, and the next attribute selected was web_traffic (accuracy = 48%). We repeated the removal process one more time, and request_URL was selected (accuracy = 63%).

For a visualization of this process, we produced stacked bar graphs in Tableau (Fig. 3) below to show how and why *SSLfinal_State* was the choice for OneR. Instead of choosing three other random attributes to compare to, we chose these next three choices for OneR after we removed the attributes as described above. It is evident that SSLfinal_State is the most predictive of the class attribute *Result.* While URL of Anchor seems to have better associations for phishing and legitimate, many of the suspicious instances are classified as legitimate, which weakens this attributes predictiveness. Overall, this chart shows an explicit visualization of the comparison of attribute predictiveness for OneR.

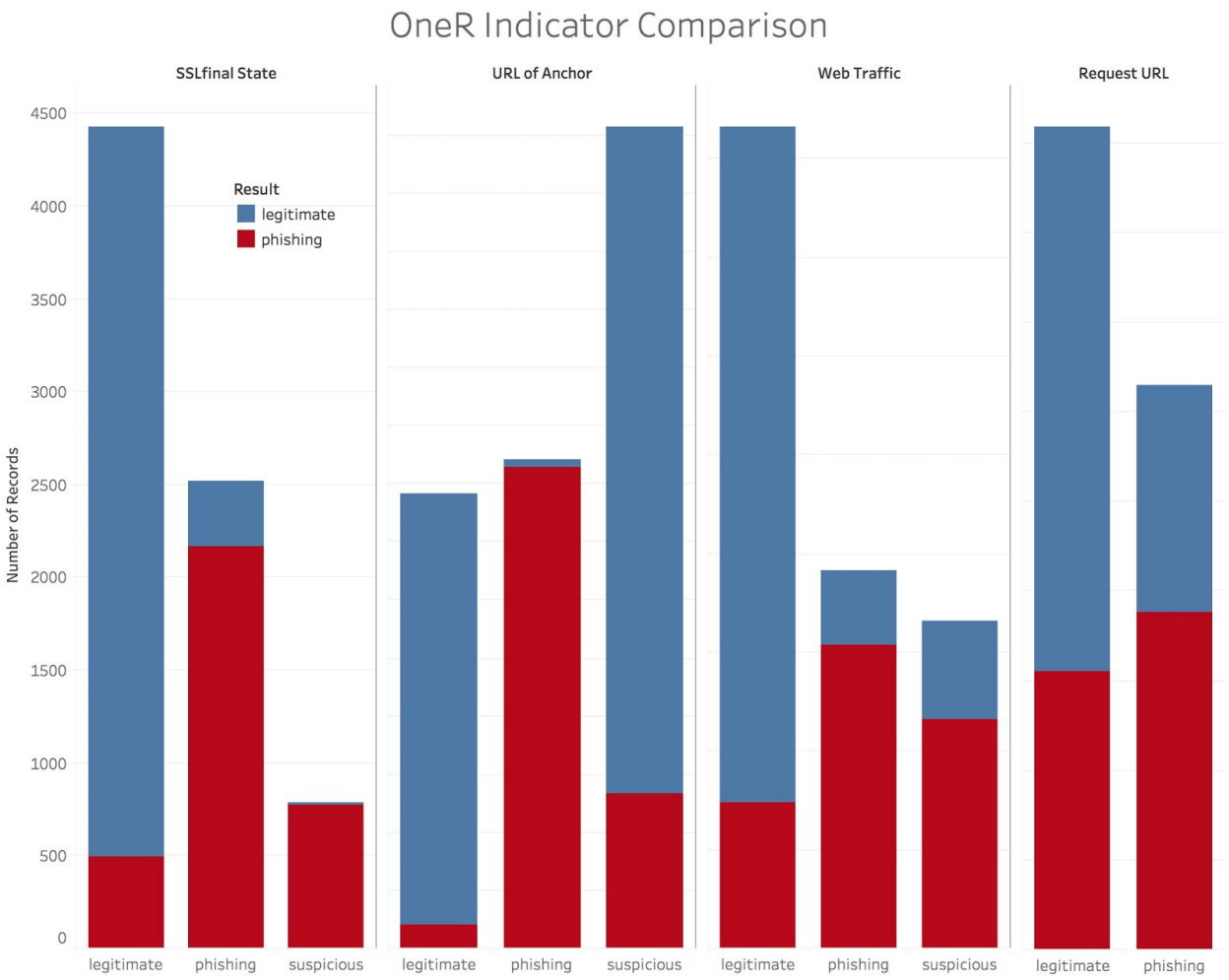

Figure 3. A stacked column comparison of predictive attributes from Tableau.

Naive Bayes

P(C), known as the a priori probability, is the probability distribution that represents the quantity of class before any evidence is offered to back the claim. In other words, a conclusion is made based on deductive reasoning and examination of existing information. Listed below are the a priori probabilities for our data set.

Class phishing: P(C) = 0.4438703
Class legitimate: P(C) = 0.5561297

The overall accuracy for this algorithm was high, as we expected it to be. Also, the accuracy of correctly predicting a phishing URL is **85.4%**.

	Predicted phishing	Predicted legitimate
Actual Phishing	1249	214
Actual Legitimate	82	1771

Confusion Matrix (trained using full training set (70%), tested using full test set (30%))

PRISM

PRISM had a low overall accuracy of 88.2388% on the test data. However, the accuracy of correctly classified phishing URLs is **92%**. While this is the highest of all of the algorithms, based on the rules created, we have determined this as overfitting.

	Predicted phishing	Predicted legitimate
Actual Phishing	1321	109
Actual Legitimate	204	1605

Confusion Matrix (trained using full training set (70%), tested using full test set (30%))

IBk

IBk, or nearest neighbor, where k = 1 had the overall highest accuracy when testing on the test data. As we increased k in the IBk algorithm, the overall accuracy decreased. Thus, we have provided the confusion matrix for only k=1. From this confusion matrix, we can tell that 203 URLs were still predicted as legitimate when they were phishing. This false negative has a higher consequence than the 83 legitimate URLs that were predicted phishing. With this algorithm, the accuracy of successfully predicting phishing is **86.1%**.

	Predicted phishing	Predicted legitimate
Actual Phishing	1260	203
Actual Legitimate	83	1770

Confusion Matrix (trained using full training set (70%), tested using full test set (30%))

J48

J48, a tree algorithm, performed with a relatively high overall accuracy. Of the non-lazy classification algorithms, it is the highest. A significant finding here is that it predicts 162 false negatives, which is far less than any other algorithm. Since false negatives are the most consequential, this is a noteworthy analysis. In this algorithm,

the accuracy of correctly predicting phishing is **89%**, which is the highest of all of the algorithms without clear overfitting. **Based off of this and laster results, we concluded that J48 is the best indicator for this project.** *See Appendix A for the tree results.*

	Predicted phishing	**Predicted legitimate**
Actual Phishing	1301	162
Actual Legitimate	127	1726

Confusion Matrix (trained using full training set (70%), tested using full test set (30%))

Confusion Matrix Discussion

False positives vs. false negatives

False negatives are much more dangerous for our phishing data set than false positives. Predicting that a URL is legitimate when it is in fact malicious is much more dangerous than predicting that a URL is malicious when it is actually safe.

In all of the confusion matrices above, the **false negatives** are located in the **upper right quadrant** of the table. For every matrix, the false negatives outnumber the false positives. This would need to be improved upon in future works if the models were to be more trusted, however we are content with the level of error encountered for this particular study due to the high accuracy levels all of our models achieved.

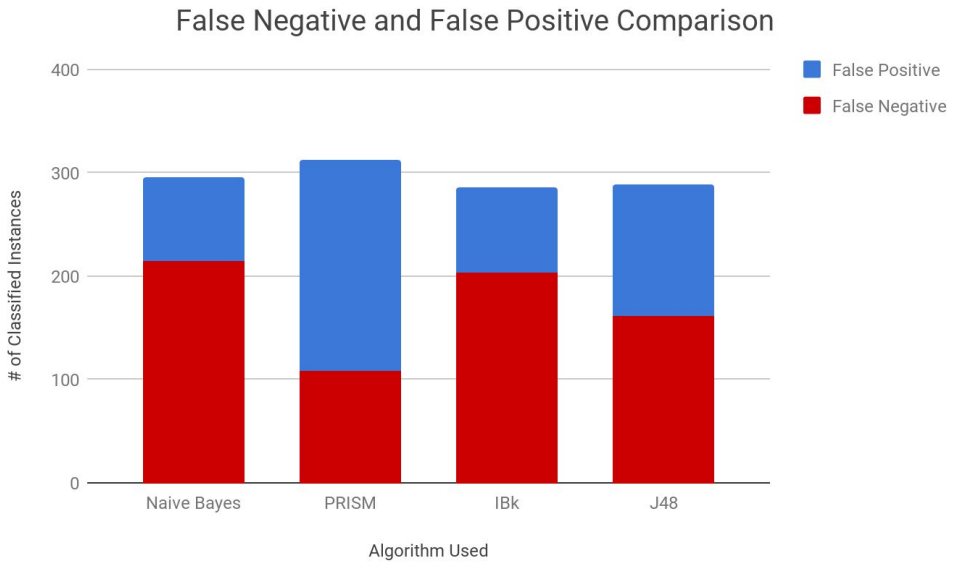

Figure 4. A stacked column comparison of false positives and false negatives for the different algorithms.

The chart above (Fig. 4) shows the comparison of false positives and false negatives of four of the five algorithms we used. We did not include OneR since that was just our baseline for the overall accuracy. It is visible here that PRISM has the least amount of false negatives, however, it has a glaring number of false positives. Based off of the rules created from PRISM, we determined that this algorithm was an example of overfitting, thus we did not rule it as the best indicator. From the chart, you can see that J48 has the second lowest number of false negatives, which is a significant difference from the other two remaining algorithms. **Thus, based on this factor, we decided J48 is the best indicator for this data.**

Application

On October 30th and November 15th, Furman students and staff were targeted by a phishing attack. The attackers spoofed the email addresses of former students and staff and attempted to get them to fill out a form with personal and login information. The links in the emails were hidden with <a> tags and they didn't display a summary or pop-up "on-hover." These were the respective <a> tags in the emails:

1. click here
2. CLICK HERE

We converted each of these "links" into instances in our data set to see if our model was able to predict that these were "phishing" emails. We had to analyze the URL's themselves and the Javascript and HTML features they took advantage of, as well as look up the URL's on domain name registration sites and in the GoDaddy WHOIS database to successfully convert these into accurate instances. If our model can successfully predict the class of these very "real world" instances, it validates our model and proves that it has useful real-world applications.

The instances, once parsed and converted, looked like this:

1.

legitimate, legitimate, legitimate, legitimate, legitimate, phishing, phishing, phishing, legitimate, phishing, phishing, legitimate,phishing,suspicious, suspicious, phishing, legitimate, legitimate,legitimate, phishing, phishing, phishing,legitimate,legitimate,phishing,phishing,phishing,phishing,phishing, legitimate, phishing

2. legitimate, legitimate, legitimate, legitimate, legitimate, phishing, phishing, phishing, phishing, phishing, phishing,legitimate, phishing,phishing, phishing, phishing, legitimate, legitimate, legitimate, phishing, phishing, phishing, legitimate, phishing, phishing, phishing,phishing, phishing, phishing, legitimate, phishing

At this point in time both sites have been taken down or blocked, so some of the information needed to perfect the instances is no longer available. These are the best approximations of the state of the sites when they were sent out.

For example, information about if the site used “iFrames” is no longer available and is not something we wrote down while the sites were still up. “iFrame” is set to legitimate in both cases based on contextual data and information about the rest of the way they implemented their attack.

Here are the key results from testing our models using these real-world instances:

J48: Correctly classifies both instances.
This is relatively unsurprising. The tree model J48 generated on our training set was the most accurate model that did not appear to be overfit. It also gave far fewer false negatives than other models.

IBk: Correctly classifies both instances with k = 1.
Based on these results, it is safe to assume that somewhere in our 10,000 instance dataset there are instances that very closely resemble our new “Furman phishing” instances. The sheer number of instances in our dataset allow Nearest Neighbor to be very effective.

PRISM: Correctly classifies both instances.
The accuracy of this model is very high, so it is not shocking that the instances were classified correctly; however, the number of rules generated by PRISM for our training set is absurdly large, which gives the impression that it is probably overfitting to the

set. Thankfully, these test instances were not affected by this likely overfitting and were still classified correctly.

Naive Bayes: Correctly classifies both instances.
With over 90% accuracy, the correct classification of our “real world” instances is exactly what we expected from the Naive Bayes model.

6. CONCLUSION

We hope that this project has demonstrated the effectiveness of data mining for creating real, usable models that can positively impact the world and have useful applications. Phishing attacks are not likely to go away anytime soon. Using this model (or ones like it) in combination with semantic analysis, antivirus software, and other protections can effectively combat the very real problem of email phishing in a smart, automated way. Email services already have the capability to check for viruses and block suspicious attachments, etc. However, it would be a great additional feature if they could analyze links/web pages in real time and make decisions about how safe they are to click on. In the attack on Furman this semester, such a feature would have prevented many students from giving up private information and losing their passwords. The need for this kind of security is clear. This project turns this conceptual security measure into a very real possibility using data mining techniques and analyses.

7. APPENDIX A

J48 Decision Tree Output

Number of Leaves : 244

Size of the tree: 424

```
SSLfinal_State = phishing
|  Prefix_Suffix = phishing
|  |  URL_of_Anchor = phishing: phishing (1455.0)
|  |  URL_of_Anchor = suspicious
|  |  |  URL_Length = legitimate
|  |  |  |  Domain_registeration_length = phishing
|  |  |  |  |  double_slash_redirecting = phishing: phishing (9.0)
|  |  |  |  |  double_slash_redirecting = legitimate
|  |  |  |  |  |  web_traffic = phishing: phishing (4.0)
|  |  |  |  |  |  web_traffic = suspicious
|  |  |  |  |  |  |  HTTPS_token = phishing: legitimate (6.0)
|  |  |  |  |  |  |  HTTPS_token = legitimate: phishing (7.0)
|  |  |  |  |  |  web_traffic = legitimate
|  |  |  |  |  |  |  having_Sub_Domain = phishing: legitimate (14.0/3.0)
|  |  |  |  |  |  |  having_Sub_Domain = suspicious: legitimate (25.0)
|  |  |  |  |  |  |  having_Sub_Domain = legitimate: phishing (1.0)
|  |  |  |  Domain_registeration_length = legitimate: phishing (14.0)
|  |  |  URL_Length = suspicious: phishing (20.0)
|  |  |  URL_Length = phishing
|  |  |  |  Shortining_Service = legitimate
|  |  |  |  |  double_slash_redirecting = phishing: phishing (21.0)
|  |  |  |  |  double_slash_redirecting = legitimate
|  |  |  |  |  |  Links_pointing_to_page = legitimate
|  |  |  |  |  |  |  HTTPS_token = phishing
|  |  |  |  |  |  |  |  Favicon = legitimate: legitimate (6.0)
|  |  |  |  |  |  |  |  Favicon = phishing: phishing (2.0)
|  |  |  |  |  |  |  HTTPS_token = legitimate
|  |  |  |  |  |  |  |  Submitting_to_email = phishing
|  |  |  |  |  |  |  |  |  Domain_registeration_length = phishing: legitimate (6.0)
|  |  |  |  |  |  |  |  |  Domain_registeration_length = legitimate: phishing (4.0)
|  |  |  |  |  |  |  |  Submitting_to_email = legitimate
|  |  |  |  |  |  |  |  |  Favicon = legitimate
|  |  |  |  |  |  |  |  |  |  popUpWindow = legitimate
|  |  |  |  |  |  |  |  |  |  |  web_traffic = phishing
|  |  |  |  |  |  |  |  |  |  |  |  having_Sub_Domain = phishing: phishing (22.0)
|  |  |  |  |  |  |  |  |  |  |  |  having_Sub_Domain = suspicious: phishing (11.0)
|  |  |  |  |  |  |  |  |  |  |  |  having_Sub_Domain = legitimate
|  |  |  |  |  |  |  |  |  |  |  |  |  Request_URL = legitimate: legitimate (13.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Request_URL = phishing: phishing (2.0)
|  |  |  |  |  |  |  |  |  |  |  web_traffic = suspicious: phishing (40.0)
|  |  |  |  |  |  |  |  |  |  |  web_traffic = legitimate
|  |  |  |  |  |  |  |  |  |  |  |  SFH = phishing
|  |  |  |  |  |  |  |  |  |  |  |  |  Links_in_tags = legitimate: phishing (21.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Links_in_tags = phishing
|  |  |  |  |  |  |  |  |  |  |  |  |  |  Request_URL = legitimate: phishing (9.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  Request_URL = phishing
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Domain_registeration_length = phishing: phishing (6.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Domain_registeration_length = legitimate
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Redirect = suspicious
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Google_Index = legitimate
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  having_At_Symbol = legitimate
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Page_Rank = phishing
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  having_Sub_Domain = phishing
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  age_of_domain = phishing: phishing (8.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  age_of_domain = legitimate: legitimate (5.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  having_Sub_Domain = suspicious: legitimate (19.0/7.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  having_Sub_Domain = legitimate: legitimate (8.0/3.0)
```

| | | | | | | | | | | | | | | | | | | Page_Rank = legitimate: phishing (6.0/1.0)
| | | | | | | | | | | | | | | | | | having_At_Symbol = phishing: legitimate (2.0)
| | | | | | | | | | | | | | | | | Google_Index = phishing: phishing (3.0)
| | | | | | | | | | | | | | | | Redirect = legitimate: phishing (3.0)
| | | | | | | | | | | | | Links_in_tags = suspicious
| | | | | | | | | | | | | | Domain_registeration_length = phishing: legitimate (5.0)
| | | | | | | | | | | | | | Domain_registeration_length = legitimate
| | | | | | | | | | | | | | | age_of_domain = phishing
| | | | | | | | | | | | | | | | Page_Rank = phishing
| | | | | | | | | | | | | | | | | Request_URL = legitimate
| | | | | | | | | | | | | | | | | | Google_Index = legitimate: phishing (7.0/1.0)
| | | | | | | | | | | | | | | | | | Google_Index = phishing: legitimate (5.0/1.0)
| | | | | | | | | | | | | | | | | Request_URL = phishing: legitimate (8.0/1.0)
| | | | | | | | | | | | | | | | Page_Rank = legitimate
| | | | | | | | | | | | | | | | | having_Sub_Domain = phishing: legitimate (3.0/1.0)
| | | | | | | | | | | | | | | | | having_Sub_Domain = suspicious: phishing (6.0)
| | | | | | | | | | | | | | | | | having_Sub_Domain = legitimate: phishing (2.0)
| | | | | | | | | | | | | | | age_of_domain = legitimate
| | | | | | | | | | | | | | | | Page_Rank = phishing
| | | | | | | | | | | | | | | | | Request_URL = legitimate: phishing (8.0)
| | | | | | | | | | | | | | | | | Request_URL = phishing
| | | | | | | | | | | | | | | | | | having_Sub_Domain = phishing
| | | | | | | | | | | | | | | | | | | Google_Index = legitimate: legitimate (9.0/4.0)
| | | | | | | | | | | | | | | | | | | Google_Index = phishing: phishing (3.0)
| | | | | | | | | | | | | | | | | | having_Sub_Domain = suspicious: phishing (0.0)
| | | | | | | | | | | | | | | | | | having_Sub_Domain = legitimate: phishing (7.0)
| | | | | | | | | | | | | | | | Page_Rank = legitimate: legitimate (4.0)
| | | | | | | | | | | | SFH = legitimate: legitimate (2.0)
| | | | | | | | | | | | SFH = suspicious: legitimate (3.0)
| | | | | | | | | | popUpWindow = phishing: legitimate (6.0)
| | | | | | | | | Favicon = phishing: phishing (6.0)
| | | | | | Links_pointing_to_page = suspicious
| | | | | | | DNSRecord = phishing: phishing (89.0)
| | | | | | | DNSRecord = legitimate
| | | | | | | | having_IP_Address = phishing: phishing (50.0)
| | | | | | | | having_IP_Address = legitimate
| | | | | | | | | Links_in_tags = legitimate
| | | | | | | | | | web_traffic = phishing: legitimate (0.0)
| | | | | | | | | | web_traffic = suspicious: phishing (2.0)
| | | | | | | | | | web_traffic = legitimate: legitimate (14.0)
| | | | | | | | | Links_in_tags = phishing
| | | | | | | | | | Favicon = legitimate
| | | | | | | | | | | Request_URL = legitimate
| | | | | | | | | | | | Page_Rank = phishing
| | | | | | | | | | | | | having_At_Symbol = legitimate
| | | | | | | | | | | | | | having_Sub_Domain = phishing
| | | | | | | | | | | | | | | Domain_registeration_length = phishing
| | | | | | | | | | | | | | | | web_traffic = phishing: phishing (3.0)
| | | | | | | | | | | | | | | | web_traffic = suspicious: legitimate (4.0)
| | | | | | | | | | | | | | | | web_traffic = legitimate: phishing (10.0)
| | | | | | | | | | | | | | | Domain_registeration_length = legitimate: phishing (5.0)
| | | | | | | | | | | | | | having_Sub_Domain = suspicious
| | | | | | | | | | | | | | | Domain_registeration_length = phishing
| | | | | | | | | | | | | | | | web_traffic = phishing: phishing (0.0)
| | | | | | | | | | | | | | | | web_traffic = suspicious: legitimate (2.0)
| | | | | | | | | | | | | | | | web_traffic = legitimate: phishing (2.0)
| | | | | | | | | | | | | | | Domain_registeration_length = legitimate: legitimate (7.0/1.0)
| | | | | | | | | | | | | | having_Sub_Domain = legitimate: phishing (3.0)
| | | | | | | | | | | | | having_At_Symbol = phishing: phishing (2.0)
| | | | | | | | | | | | Page_Rank = legitimate: phishing (19.0/2.0)
| | | | | | | | | | | Request_URL = phishing
| | | | | | | | | | | | age_of_domain = phishing
| | | | | | | | | | | | | having_Sub_Domain = phishing: phishing (38.0)
| | | | | | | | | | | | | having_Sub_Domain = suspicious
| | | | | | | | | | | | | | Page_Rank = phishing
| | | | | | | | | | | | | | | web_traffic = phishing: phishing (2.0)

| | | | | | | | | | | | | | | | | web_traffic = suspicious: phishing (8.0)
| | | | | | | | | | | | | | | | | web_traffic = legitimate
| | | | | | | | | | | | | | | | | | Domain_registeration_length = phishing: phishing (2.0)
| | | | | | | | | | | | | | | | | | Domain_registeration_length = legitimate: legitimate (2.0)
| | | | | | | | | | | | | | | | | Page_Rank = legitimate: phishing (17.0)
| | | | | | | | | | | | | | | | having_Sub_Domain = legitimate: phishing (6.0)
| | | | | | | | | | | | | | | age_of_domain = legitimate
| | | | | | | | | | | | | | | | web_traffic = phishing: phishing (2.0)
| | | | | | | | | | | | | | | | web_traffic = suspicious: phishing (4.0)
| | | | | | | | | | | | | | | | web_traffic = legitimate
| | | | | | | | | | | | | | | | | having_Sub_Domain = phishing: legitimate (5.0/1.0)
| | | | | | | | | | | | | | | | | having_Sub_Domain = suspicious: legitimate (0.0)
| | | | | | | | | | | | | | | | | having_Sub_Domain = legitimate: phishing (2.0)
| | | | | | | | | | | | | | | Favicon = phishing: legitimate (6.0)
| | | | | | | | | | | | | | | Links_in_tags = suspicious
| | | | | | | | | | | | | | | Google_Index = legitimate
| | | | | | | | | | | | | | | | Favicon = legitimate
| | | | | | | | | | | | | | | | age_of_domain = phishing
| | | | | | | | | | | | | | | | | Statistical_report = phishing: phishing (2.0)
| | | | | | | | | | | | | | | | | Statistical_report = legitimate
| | | | | | | | | | | | | | | | | | Page_Rank = phishing
| | | | | | | | | | | | | | | | | | having_Sub_Domain = phishing
| | | | | | | | | | | | | | | | | | | Domain_registeration_length = phishing: phishing (11.0/2.0)
| | | | | | | | | | | | | | | | | | | Domain_registeration_length = legitimate
| | | | | | | | | | | | | | | | | | | | Request_URL = legitimate: phishing (3.0)
| | | | | | | | | | | | | | | | | | | | Request_URL = phishing: legitimate (7.0/2.0)
| | | | | | | | | | | | | | | | | | having_Sub_Domain = suspicious: legitimate (8.0/1.0)
| | | | | | | | | | | | | | | | | | having_Sub_Domain = legitimate: phishing (2.0)
| | | | | | | | | | | | | | | | | Page_Rank = legitimate
| | | | | | | | | | | | | | | | | | having_Sub_Domain = phishing: phishing (3.0)
| | | | | | | | | | | | | | | | | | having_Sub_Domain = suspicious: phishing (14.0)
| | | | | | | | | | | | | | | | | | having_Sub_Domain = legitimate: legitimate (3.0/1.0)
| | | | | | | | | | | | | | | | age_of_domain = legitimate
| | | | | | | | | | | | | | | | | having_Sub_Domain = phishing: phishing (2.0)
| | | | | | | | | | | | | | | | | having_Sub_Domain = suspicious: legitimate (4.0)
| | | | | | | | | | | | | | | | | having_Sub_Domain = legitimate: legitimate (4.0)
| | | | | | | | | | | | | | | Favicon = phishing: phishing (5.0)
| | | | | | | | | | | | | | | Google_Index = phishing: phishing (33.0)
| | | | | | | | | | | | | | | Links_pointing_to_page = phishing: phishing (18.0)
| | | | | | | | | | | | | | Shortining_Service = phishing
| | | | | | | | | | | | | | | Links_in_tags = legitimate: phishing (1.0)
| | | | | | | | | | | | | | | Links_in_tags = phishing: phishing (6.0)
| | | | | | | | | | | | | | | Links_in_tags = suspicious: legitimate (14.0)
| | | | | | | | | | | | | | URL_of_Anchor = legitimate
| | | | | | | | | | | | | | | Domain_registeration_length = phishing
| | | | | | | | | | | | | | | RightClick = legitimate
| | | | | | | | | | | | | | | | Google_Index = legitimate
| | | | | | | | | | | | | | | | | web_traffic = phishing
| | | | | | | | | | | | | | | | | | Page_Rank = phishing: legitimate (7.0/1.0)
| | | | | | | | | | | | | | | | | | Page_Rank = legitimate: phishing (6.0)
| | | | | | | | | | | | | | | | | web_traffic = suspicious: phishing (4.0)
| | | | | | | | | | | | | | | | | web_traffic = legitimate
| | | | | | | | | | | | | | | | | | Redirect = suspicious: legitimate (57.0)
| | | | | | | | | | | | | | | | | | Redirect = legitimate
| | | | | | | | | | | | | | | | | | | Shortining_Service = legitimate: phishing (4.0)
| | | | | | | | | | | | | | | | | | | Shortining_Service = phishing: legitimate (5.0)
| | | | | | | | | | | | | | | | | Google_Index = phishing
| | | | | | | | | | | | | | | | | | having_IP_Address = phishing: legitimate (4.0/1.0)
| | | | | | | | | | | | | | | | | | having_IP_Address = legitimate: phishing (13.0)
| | | | | | | | | | | | | | | | RightClick = phishing: phishing (12.0)
| | | | | | | | | | | | | | | Domain_registeration_length = legitimate: phishing (34.0)
| | | | | | | | | | | | | | Prefix_Suffix = legitimate: legitimate (74.0)
SSLfinal_State = legitimate
| | | | | | | | | | | | | | web_traffic = phishing
| | | | | | | | | | | | | | | URL_of_Anchor = phishing: phishing (53.0)
| | | | | | | | | | | | | | | URL_of_Anchor = suspicious: legitimate (210.0/2.0)

| | URL_of_Anchor = legitimate
| | | Statistical_report = phishing
| | | | Links_in_tags = legitimate: legitimate (4.0)
| | | | Links_in_tags = phishing: phishing (1.0)
| | | | Links_in_tags = suspicious: legitimate (4.0)
| | | Statistical_report = legitimate: legitimate (133.0/1.0)
| web_traffic = suspicious
| | URL_of_Anchor = phishing: phishing (84.0)
| | URL_of_Anchor = suspicious
| | | Prefix_Suffix = phishing
| | | | RightClick = legitimate
| | | | | having_Sub_Domain = phishing
| | | | | | Request_URL = legitimate
| | | | | | | Page_Rank = phishing
| | | | | | | | Submitting_to_email = phishing: phishing (4.0)
| | | | | | | | Submitting_to_email = legitimate
| | | | | | | | | on_mouseover = legitimate
| | | | | | | | | | Abnormal_URL = phishing: legitimate (4.0)
| | | | | | | | | | Abnormal_URL = legitimate
| | | | | | | | | | | having_IP_Address = phishing: phishing (4.0)
| | | | | | | | | | | having_IP_Address = legitimate
| | | | | | | | | | | | Links_pointing_to_page = legitimate
| | | | | | | | | | | | | Links_in_tags = legitimate: phishing (4.0)
| | | | | | | | | | | | | Links_in_tags = phishing: legitimate (9.0/1.0)
| | | | | | | | | | | | | Links_in_tags = suspicious: legitimate (11.0/3.0)
| | | | | | | | | | | | Links_pointing_to_page = suspicious
| | | | | | | | | | | | | DNSRecord = phishing
| | | | | | | | | | | | | | Favicon = legitimate: phishing (16.0)
| | | | | | | | | | | | | | Favicon = phishing
| | | | | | | | | | | | | | | Links_in_tags = legitimate: legitimate (2.0)
| | | | | | | | | | | | | | | Links_in_tags = phishing: phishing (0.0)
| | | | | | | | | | | | | | | Links_in_tags = suspicious: phishing (2.0)
| | | | | | | | | | | | | DNSRecord = legitimate
| | | | | | | | | | | | | | Shortining_Service = legitimate
| | | | | | | | | | | | | | | age_of_domain = phishing
| | | | | | | | | | | | | | | | Google_Index = legitimate
| | | | | | | | | | | | | | | | | Links_in_tags = legitimate: legitimate (10.0/4.0)
| | | | | | | | | | | | | | | | | Links_in_tags = phishing: phishing (10.0/4.0)
| | | | | | | | | | | | | | | | | Links_in_tags = suspicious: legitimate (7.0/2.0)
| | | | | | | | | | | | | | | | Google_Index = phishing: phishing (6.0)
| | | | | | | | | | | | | | | age_of_domain = legitimate
| | | | | | | | | | | | | | | | Google_Index = legitimate
| | | | | | | | | | | | | | | | | Links_in_tags = legitimate: legitimate (2.0)
| | | | | | | | | | | | | | | | | Links_in_tags = phishing: phishing (6.0/2.0)
| | | | | | | | | | | | | | | | | Links_in_tags = suspicious: legitimate (8.0/2.0)
| | | | | | | | | | | | | | | | Google_Index = phishing: legitimate (6.0)
| | | | | | | | | | | | | | | Shortining_Service = phishing: legitimate (4.0)
| | | | | | | | | | | | | | Links_pointing_to_page = phishing: phishing (4.0)
| | | | | | | | | | | | | on_mouseover = phishing: phishing (3.0)
| | | | | | | | | | | | Page_Rank = legitimate: legitimate (11.0/2.0)
| | | | | | | | | | | Request_URL = phishing
| | | | | | | | | | | | Shortining_Service = legitimate: phishing (42.0/3.0)
| | | | | | | | | | | | Shortining_Service = phishing: legitimate (3.0)
| | | | | | | | | | | having_Sub_Domain = suspicious
| | | | | | | | | | | | Links_in_tags = legitimate: phishing (15.0)
| | | | | | | | | | | | Links_in_tags = phishing
| | | | | | | | | | | | | Domain_registeration_length = phishing
| | | | | | | | | | | | | | Page_Rank = phishing
| | | | | | | | | | | | | | | Request_URL = legitimate
| | | | | | | | | | | | | | | | Links_pointing_to_page = legitimate
| | | | | | | | | | | | | | | | | URL_Length = legitimate: phishing (2.0)
| | | | | | | | | | | | | | | | | URL_Length = suspicious: phishing (0.0)
| | | | | | | | | | | | | | | | | URL_Length = phishing: legitimate (2.0)
| | | | | | | | | | | | | | | | Links_pointing_to_page = suspicious: phishing (6.0)
| | | | | | | | | | | | | | | | Links_pointing_to_page = phishing: phishing (0.0)
| | | | | | | | | | | | | | | Request_URL = phishing: legitimate (7.0/1.0)

| | | | | | | | Page_Rank = legitimate: phishing (15.0/1.0)
| | | | | | | Domain_registeration_length = legitimate: phishing (18.0)
| | | | | | Links_in_tags = suspicious
| | | | | | | Abnormal_URL = phishing: legitimate (6.0)
| | | | | | | Abnormal_URL = legitimate
| | | | | | | | age_of_domain = phishing
| | | | | | | | | Shortining_Service = legitimate
| | | | | | | | | | Request_URL = legitimate
| | | | | | | | | | | DNSRecord = phishing: phishing (6.0)
| | | | | | | | | | | DNSRecord = legitimate
| | | | | | | | | | | | having_IP_Address = phishing: phishing (11.0/2.0)
| | | | | | | | | | | | having_IP_Address = legitimate
| | | | | | | | | | | | | Links_pointing_to_page = legitimate: legitimate (1.0)
| | | | | | | | | | | | | Links_pointing_to_page = suspicious: phishing (6.0/2.0)
| | | | | | | | | | | | | Links_pointing_to_page = phishing: legitimate (2.0)
| | | | | | | | | | Request_URL = phishing: legitimate (8.0/1.0)
| | | | | | | | | Shortining_Service = phishing: legitimate (2.0)
| | | | | | | | age_of_domain = legitimate: legitimate (4.0)
| | | | | having_Sub_Domain = legitimate
| | | | | | Favicon = legitimate
| | | | | | | Statistical_report = phishing
| | | | | | | | Links_in_tags = legitimate: phishing (5.0)
| | | | | | | | Links_in_tags = phishing: phishing (9.0/2.0)
| | | | | | | | Links_in_tags = suspicious: legitimate (2.0)
| | | | | | | Statistical_report = legitimate
| | | | | | | | Submitting_to_email = phishing: legitimate (8.0)
| | | | | | | | Submitting_to_email = legitimate
| | | | | | | | | SFH = phishing
| | | | | | | | | | Links_pointing_to_page = legitimate
| | | | | | | | | | | age_of_domain = phishing: legitimate (22.0/1.0)
| | | | | | | | | | | age_of_domain = legitimate
| | | | | | | | | | | | Google_Index = legitimate
| | | | | | | | | | | | | having_IP_Address = phishing
| | | | | | | | | | | | | | Domain_registeration_length = phishing: phishing (3.0/1.0)
| | | | | | | | | | | | | | Domain_registeration_length = legitimate: legitimate (2.0)
| | | | | | | | | | | | | having_IP_Address = legitimate: legitimate (19.0/2.0)
| | | | | | | | | | | | Google_Index = phishing
| | | | | | | | | | | | | Links_in_tags = legitimate: phishing (0.0)
| | | | | | | | | | | | | Links_in_tags = phishing: legitimate (5.0/1.0)
| | | | | | | | | | | | | Links_in_tags = suspicious: phishing (3.0)
| | | | | | | | | | Links_pointing_to_page = suspicious
| | | | | | | | | | | Links_in_tags = legitimate: phishing (13.0)
| | | | | | | | | | | Links_in_tags = phishing
| | | | | | | | | | | | DNSRecord = phishing: phishing (11.0)
| | | | | | | | | | | | DNSRecord = legitimate
| | | | | | | | | | | | | HTTPS_token = phishing: phishing (3.0)
| | | | | | | | | | | | | HTTPS_token = legitimate
| | | | | | | | | | | | | | age_of_domain = phishing
| | | | | | | | | | | | | | | Page_Rank = phishing
| | | | | | | | | | | | | | | | Redirect = suspicious: phishing (9.0/1.0)
| | | | | | | | | | | | | | | | Redirect = legitimate: legitimate (2.0)
| | | | | | | | | | | | | | | Page_Rank = legitimate
| | | | | | | | | | | | | | | | Redirect = suspicious: legitimate (8.0/2.0)
| | | | | | | | | | | | | | | | Redirect = legitimate: phishing (2.0)
| | | | | | | | | | | | | | age_of_domain = legitimate: legitimate (15.0/3.0)
| | | | | | | | | | Links_in_tags = suspicious
| | | | | | | | | | | DNSRecord = phishing
| | | | | | | | | | | | Domain_registeration_length = phishing
| | | | | | | | | | | | | age_of_domain = phishing: legitimate (2.0)
| | | | | | | | | | | | | age_of_domain = legitimate: phishing (7.0/2.0)
| | | | | | | | | | | | Domain_registeration_length = legitimate: legitimate (2.0)
| | | | | | | | | | | DNSRecord = legitimate: legitimate (30.0/3.0)
| | | | | | | | | Links_pointing_to_page = phishing
| | | | | | | | | | DNSRecord = phishing: phishing (2.0)
| | | | | | | | | | DNSRecord = legitimate: legitimate (12.0/1.0)
| | | | | | | | SFH = legitimate

| | | | | | | | | | DNSRecord = phishing: phishing (12.0/1.0)
| | | | | | | | | | DNSRecord = legitimate: legitimate (4.0)
| | | | | | | | | SFH = suspicious: legitimate (2.0)
| | | | | | Favicon = phishing
| | | | | | | Submitting_to_email = phishing
| | | | | | | | Links_in_tags = legitimate: phishing (3.0)
| | | | | | | | Links_in_tags = phishing: legitimate (4.0)
| | | | | | | | Links_in_tags = suspicious: legitimate (0.0)
| | | | | | | Submitting_to_email = legitimate: legitimate (36.0)
| | | | RightClick = phishing: legitimate (18.0)
| | | Prefix_Suffix = legitimate: legitimate (37.0)
| | URL_of_Anchor = legitimate: legitimate (120.0/1.0)
| web_traffic = legitimate
| | Links_in_tags = legitimate
| | | Iframe = legitimate: legitimate (954.0)
| | | Iframe = phishing
| | | | Domain_registeration_length = phishing: legitimate (85.0)
| | | | Domain_registeration_length = legitimate
| | | | | Prefix_Suffix = phishing
| | | | | | URL_of_Anchor = phishing: legitimate (0.0)
| | | | | | URL_of_Anchor = suspicious: phishing (4.0)
| | | | | | URL_of_Anchor = legitimate: legitimate (8.0)
| | | | | Prefix_Suffix = legitimate: legitimate (14.0)
| | Links_in_tags = phishing
| | | URL_of_Anchor = phishing
| | | | Prefix_Suffix = phishing: phishing (10.0)
| | | | Prefix_Suffix = legitimate: legitimate (6.0)
| | | URL_of_Anchor = suspicious
| | | | Request_URL = legitimate
| | | | | Favicon = legitimate
| | | | | | Page_Rank = phishing
| | | | | | | Shortining_Service = legitimate
| | | | | | | | Google_Index = legitimate
| | | | | | | | | Prefix_Suffix = phishing
| | | | | | | | | | SFH = phishing
| | | | | | | | | | | having_Sub_Domain = phishing: legitimate (36.0/6.0)
| | | | | | | | | | | having_Sub_Domain = suspicious
| | | | | | | | | | | | Domain_registeration_length = phishing
| | | | | | | | | | | | | Links_pointing_to_page = legitimate: phishing (9.0/4.0)
| | | | | | | | | | | | | Links_pointing_to_page = suspicious: legitimate (6.0/1.0)
| | | | | | | | | | | | | Links_pointing_to_page = phishing: legitimate (4.0)
| | | | | | | | | | | | Domain_registeration_length = legitimate: phishing (3.0)
| | | | | | | | | | | having_Sub_Domain = legitimate: legitimate (18.0/2.0)
| | | | | | | | | | SFH = legitimate: legitimate (15.0)
| | | | | | | | | | SFH = suspicious: legitimate (9.0)
| | | | | | | | | Prefix_Suffix = legitimate: legitimate (19.0)
| | | | | | | | Google_Index = phishing
| | | | | | | | | Links_pointing_to_page = legitimate: legitimate (2.0)
| | | | | | | | | Links_pointing_to_page = suspicious: phishing (3.0)
| | | | | | | | | Links_pointing_to_page = phishing: phishing (0.0)
| | | | | | | Shortining_Service = phishing: legitimate (47.0)
| | | | | | Page_Rank = legitimate: legitimate (78.0)
| | | | | Favicon = phishing: legitimate (127.0)
| | | | Request_URL = phishing
| | | | | age_of_domain = phishing
| | | | | | having_Sub_Domain = phishing
| | | | | | | SFH = phishing: phishing (33.0/1.0)
| | | | | | | SFH = legitimate: legitimate (3.0)
| | | | | | | SFH = suspicious: legitimate (1.0)
| | | | | | having_Sub_Domain = suspicious
| | | | | | | DNSRecord = phishing: legitimate (11.0)
| | | | | | | DNSRecord = legitimate
| | | | | | | | SFH = phishing
| | | | | | | | | having_IP_Address = phishing: phishing (6.0)
| | | | | | | | | having_IP_Address = legitimate
| | | | | | | | | Page_Rank = phishing: legitimate (5.0)

| | | | | | | | | | Page_Rank = legitimate
| | | | | | | | | | | Domain_registeration_length = phishing: legitimate (5.0/1.0)
| | | | | | | | | | | Domain_registeration_length = legitimate: phishing (3.0)
| | | | | | | | SFH = legitimate: legitimate (3.0)
| | | | | | | | SFH = suspicious: legitimate (3.0)
| | | | | | having_Sub_Domain = legitimate: legitimate (23.0)
| | | | | age_of_domain = legitimate: legitimate (66.0)
| | | URL_of_Anchor = legitimate: legitimate (266.0/1.0)
| | Links_in_tags = suspicious
| | | Request_URL = legitimate: legitimate (1010.0)
| | | Request_URL = phishing
| | | | URL_of_Anchor = phishing: phishing (1.0)
| | | | URL_of_Anchor = suspicious
| | | | | DNSRecord = phishing
| | | | | | Links_pointing_to_page = legitimate: legitimate (41.0)
| | | | | | Links_pointing_to_page = suspicious
| | | | | | | having_Sub_Domain = phishing
| | | | | | | | having_IP_Address = phishing: legitimate (2.0)
| | | | | | | | having_IP_Address = legitimate: phishing (3.0)
| | | | | | | having_Sub_Domain = suspicious: legitimate (5.0)
| | | | | | | having_Sub_Domain = legitimate: legitimate (10.0)
| | | | | | Links_pointing_to_page = phishing: legitimate (0.0)
| | | | | DNSRecord = legitimate: legitimate (157.0)
| | | | URL_of_Anchor = legitimate: legitimate (105.0)
SSLfinal_State = suspicious
| URL_of_Anchor = phishing: phishing (663.0)
| URL_of_Anchor = suspicious
| | URL_Length = legitimate: legitimate (9.0/1.0)
| | URL_Length = suspicious: phishing (0.0)
| | URL_Length = phishing
| | | having_Sub_Domain = phishing: phishing (20.0)
| | | having_Sub_Domain = suspicious: phishing (55.0)
| | | having_Sub_Domain = legitimate
| | | | age_of_domain = phishing: phishing (19.0/2.0)
| | | | age_of_domain = legitimate: legitimate (4.0)
| URL_of_Anchor = legitimate: phishing (18.0)