MACHINE LEARNING MODEL

COMPARISON BASED ON SOME METRICS

Safa ORHAN

Computer Engineering Student

Istanbul Kultur University

Istanbul, Turkey

Eyüp USTA

Computer Engineering Student

Istanbul Kultur University

Istanbul, Turkey

I. SUPPORT VECTOR MACHINE

Kernels = linear, poly, rbf, sigmoid

C = 1, 2, 3, 4, 5

Degree = 1, 2, 3, 4, 5, 6

Gamma = scale, auto

Decision function shape = ovo, ovr

A. Results of SVM:

Test Size: 15%

Best Accuracy = 96.44%

Best Combination = “poly, 5, 5, scale, ovo”

Worst Accuracy = 82.64%

Worst Combination = “sigmoid, 5, 1, scale, ovo”

Average Accuracy = 91.42%

Test Size: 30%

Best Accuracy = 98.10%

Best Combination = “poly, 5, 6, scale, ovo”

Worst Accuracy = 83.78%

Worst Combination = “sigmoid, 3, 1, scale, ovo”

Average Accuracy = 92.42%

Test Size: 35%

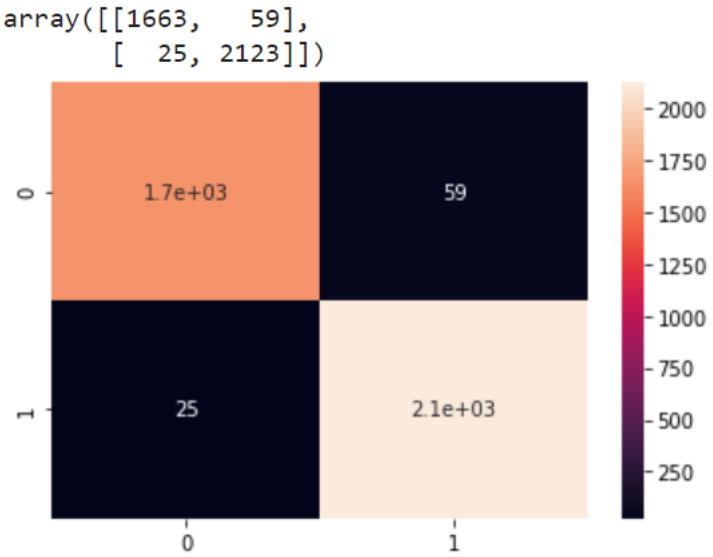
Best Accuracy = 98.03%

Best Combination = “poly, 5, 6, scale, ovo”

Worst Accuracy = 82.79%

Worst Combination = “sigmoid, 3, 1, scale, ovo”

Average Accuracy = 91.98%



II. LINEAR SUPPORT VECTOR MACHINE

Losses = hinge, squared hinge

Penalty = l2

C = 1, 2, 3, 4, 5

Multi Class = ovr, crammer singer

A. Resultf of Linear SVC:

Test Size: 15%

Best Accuracy = 92.40%

Best Combination = “squared_hinge, l2, 1, ovr”

Worst Accuracy = 92.22%

Worst Combination = “hinge, l2, 1, ovr”

Average Accuracy = 92.33%

Test Size: 30%

Best Accuracy = 93.15%

Best Combination = “hinge, l2, 5, ovr”

Worst Accuracy = 93.03%

Worst Combination = “hinge, l2, 4, ovr”

Average Accuracy = 93.08%

Test Size: 35%

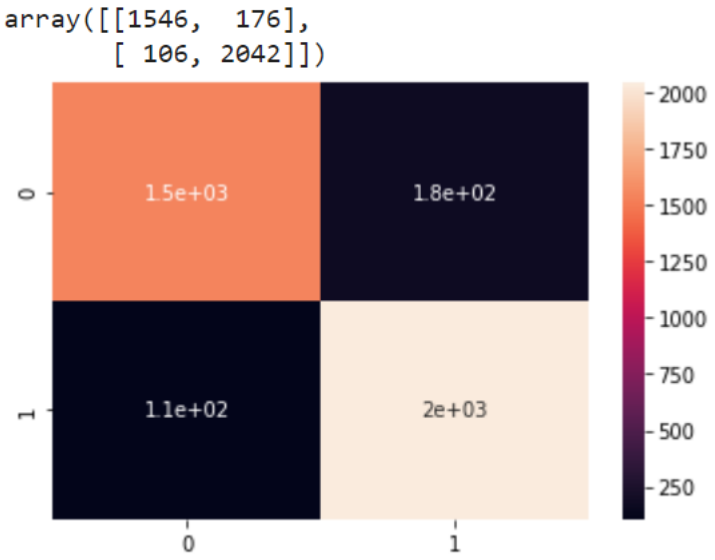
Best Accuracy = 92.97%

Best Combination = “squared_hinge, l2, 1, ovr”

Worst Accuracy = 92.79%

Worst Combination = “hinge, l2, 1, ovr”

Average Accuracy = 92.83%



III. K – NEAREST NEIGHBORS

K = 1,2,3,4,5,6,7,8,9,10

Weights = uniform, distance

Metric = euclidean, manhattan, chebyshev, minkowski

A. Results of KNN:

Test Size: 15%

Best Accuracy = 96.14%

Best Combination = “8, disctance, brute, manhattan”

Worst Accuracy = 71.18%

Worst Combination = “8, uniform, auto, chebyshev”

Average Accuracy = 91.86%

Test Size: 30%

Best Accuracy = 98.52%

Best Combination = “5, distance, auto, manhattan”

Worst Accuracy = 78.17%

Worst Combination = “8, uniform, auto, chebyshev”

Average Accuracy = 95.99%

Test Size: 35%

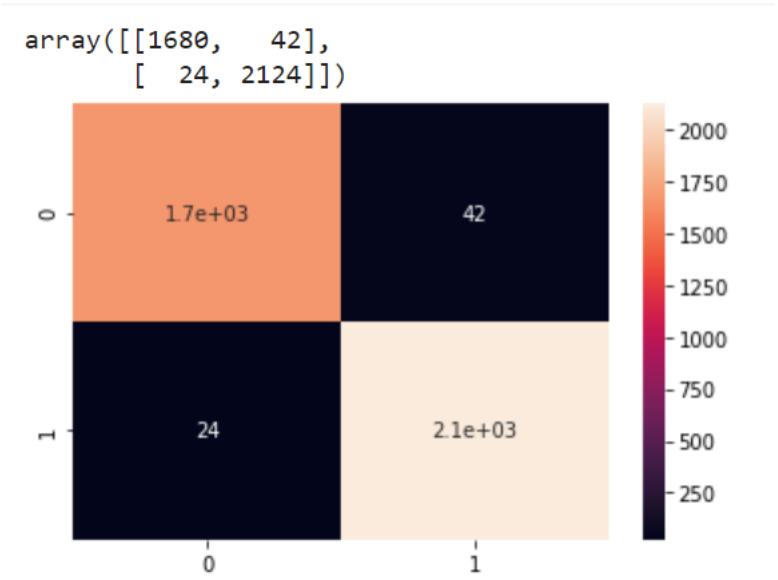
Best Accuracy = 98.44%

Best Combination = “5, distance, auto, manhattan”

Worst Accuracy = 77.08%

Worst Combination = “8, uniform, auto, chebyshev”

Average Accuracy = 95.85%



IV. DECISION TREE CLASSIFIER

Max features = None, auto, sqrt, log2

Criterion = gini, entropy

Splitter = best, random

A. Result of Decision Tree Classifier:

Test Size: 15%

Best Accuracy = 96.92%

Best Combination = “None, None, entropy, random"

Worst Accuracy = 93.91%

Worst Combination = “None, log2, gini, random”

Average Accuracy = 95.64%

Test Size: 30%

Best Accuracy = 95.96%

Best Combination = “None, None, entropy, random”

Worst Accuracy = 94.33%

Worst Combination = “None, log2, gini, random”

Average Accuracy = 95.37%

Test Size: 35%

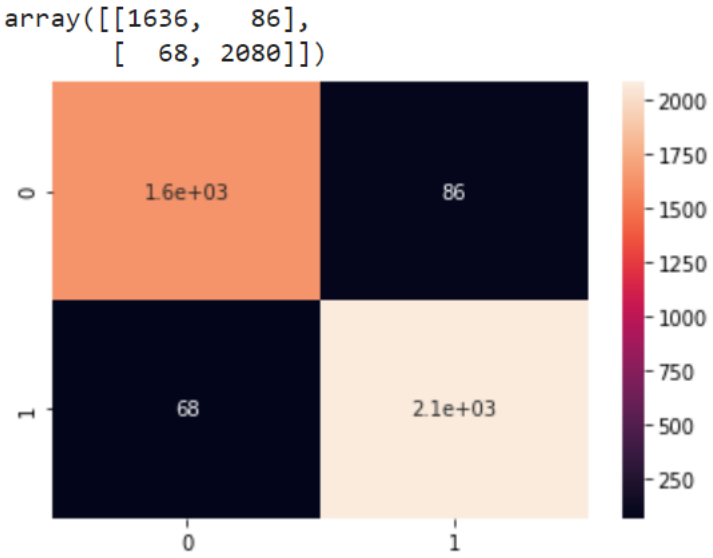
Best Accuracy = 95.86%

Best Combination = “default parameter”

Worst Accuracy = 94.03%

Worst Combination = “balanced, log2, entropy, random”

Average Accuracy = 95.18%



V. GAUSSIAN NAIVE BAYES CLASSIFIER

Test Size: 15%

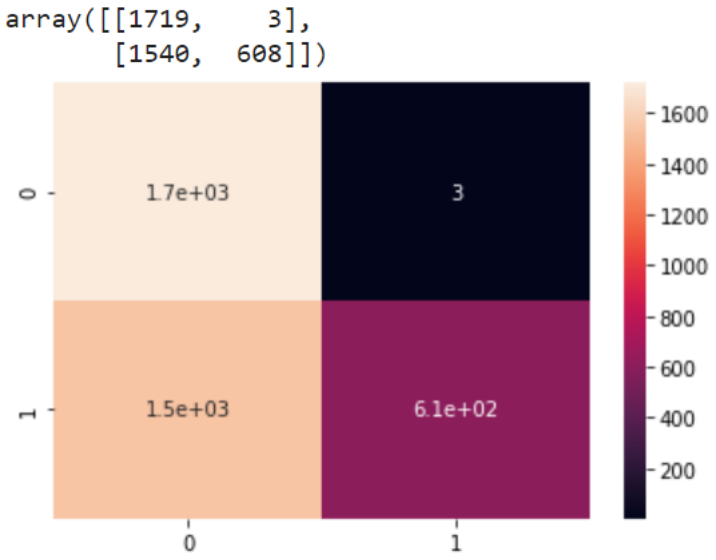
Accuracy = 59.25%

Test Size: 30%

Accuracy = 60.47%

Test Size: 35%

Accuracy = 59.94%



VI. BERNOULLI NAIVE BAYES CLASSIFIER

The accuracy score of our Naive Bayes classifier is 56.54%.

A. Bernoulli Naive Bayes Classifier:

Alpha = 0, 1, 2, 3, 4, 5, 7, 9, 11

Binarize = 0, 1, 2, 3, 4, 5, 7, 9, 11

Fit prior = True, False

B. Results of BernoulliNB:

Test Size: 15%

Best Accuracy = 89.93%

Best Combination = “0, 0, true”

Worst Accuracy = 56.60%

Worst Combination = “0 1.0 true”

Average Accuracy = 60.07%

Test Size: 35%

Best Accuracy = 90.86%

Best Combination = “0, 0, false”

Worst Accuracy = 56.01%

Worst Combination = “0, 1.0, true”

Average Accuracy = 60.07%

Test Size: 35%

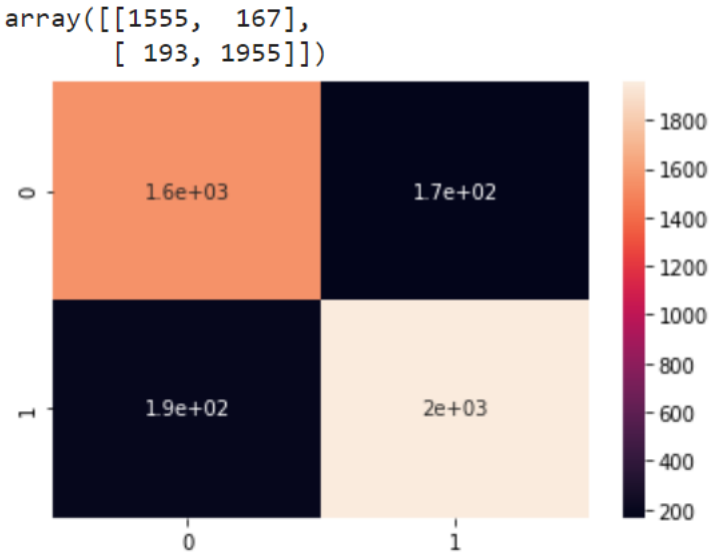
Best Accuracy = 90.08%

Best Combination = “3.0, 0, true”

Worst Accuracy = 55.50%

Worst Combination = “0, 1.0, true”

Average Accuracy = 59.61%



VII. RANDOM FOREST CLASSIFIER

Max features = None, auto, sqrt, log2

Criterion = gini, entropy

Class_weight = None, balanced, balanced_subsample

Warm start = True, False

A. Result of Random Forest Classifier:

Test Size: 15%

Best Accuracy = 97.34%

Best Combination = “None, auto, entropy, false”

Worst Accuracy = 96.68%

Worst Combination = “balanced, none, entropy, true”

Average Accuracy = 97.00%

Test Size: 30%

Best Accuracy = 96.95%

Best Combination = “None, log2, entropy, false”

Worst Accuracy = 96.50%

Worst Combination = “balanced_subsample, none, entropy, true”

Average Accuracy = 96.72%

Test Size: 35%

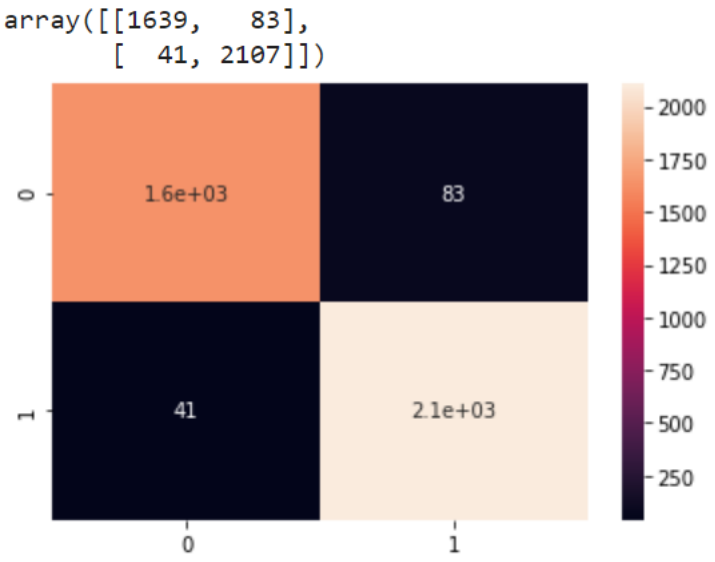
Best Accuracy = 96.89%

Best Combination = “None, log2, gini, true”

Worst Accuracy = 96.50%

Worst Combination = “None, None, gini, true”

Average Accuracy = 96.72%



VIII. DEEP LEARNING WITH TENSORFLOW

Optimizers = 'sgd', 'rmsprop', 'adam', 'adadelta', 'adagrad', 'adamax', 'nadam', 'ftrl'

Loss = 'binary_crossentropy', 'categorical_crossentropy', 'hinge', 'squared_hinge', 'huber'

Activation = 'softplus', 'softsign', 'selu', 'elu', 'exponential', 'tanh', 'sigmoid', 'relu'

A. Results of the Deep Network:

Input Layer: Dense (32, , activation = relu, input_shape = (30,2))

Deep Layers: Dense (64, activation = relu), Dense (32, activation = relu)

Output Layer: Dense (1, activation=softplus)

Test Size: 15%

Best Accuracy = 55.98%

Best Combination = “softplus, adamax, huber”

Worst Accuracy = 44.04%

Worst Combination = “softplus, sgd, huber”

Average Accuracy = 52.92%

Input Layer: Dense (64, kernel_regularizer=l2, activation = relu, input_shape = (30,2))

Deep Layers: Dense (128, kernel_regularizer=l2, activation = relu), Dense (256, kernel_regularizer=l2, activation = relu), Dense (512, kernel_regularizer=l2, activation = relu), Dropout(0.3)

Output Layer: Dense (1, activation=softplus)

Test Size: 15%

Best Accuracy = 54.74%

Best Combination = “softplus, adagrad, huber”

Worst Accuracy = 45.12%

Worst Combination = “softplus, sgd, huber”

Average Accuracy = 50.73%

Input Layer: Dense (64, activation = relu, input_shape = (30,2))

Deep Layers: Dense (128, activation = relu), Dense (128, activation = relu),

Output Layer: Dense (1, activation=softplus)

Test Size: 15%

Best Accuracy = 75.76%

Best Combination = “softplus, adamax, binary_crossentropy”

Worst Accuracy = 44.35%

Worst Combination = “softplus, sgd, huber”

Average Accuracy = 69.04%

Test Size: 35%

Best Accuracy = 75.54%

Best Combination = “softplus, adamax, binary_crossentropy”

Worst Accuracy = 44.31%

Worst Combination = “softplus, sgd, huber”

Average Accuracy = 69.19%

Input Layer: Dense (64, kernel_regularizer=l2, activation = relu, input_shape = (30,2))

Deep Layers: Dense (128, kernel_regularizer=l2, activation = relu), Dense (128, kernel_regularizer=l2, activation = relu),

Output Layer: Dense (1, activation=softplus)

Test Size: 35%

Best Accuracy = 76.20%

Best Combination = “softplus, nadam, huber”

Worst Accuracy = 44.55%

Worst Combination = “softplus, sgd, huber”

Average Accuracy = 69.62%

Input Layer: Dense (64, kernel_regularizer=l2, activation = relu, input_shape = (30,2))

Deep Layers: Dense (128, kernel_regularizer=l2, activation = relu), Dense (128, kernel_regularizer=l2, activation = relu), Dense (128, kernel_regularizer=l2, activation = relu), Dense (256, kernel_regularizer=l2, activation = relu),

Output Layer: Dense (1, activation=softplus)

Test Size: 35%

Best Accuracy = 75.40%

Best Combination = “softplus, adamx, binary_crossentropy”

Worst Accuracy = 44.34%

Worst Combination = “softplus, sgd, huber”

Average Accuracy = 68.96%

Input Layer: Dense (64, kernel_regularizer=l2, activation = relu, input_shape = (30,2))

Deep Layers: Dense (128, kernel_regularizer=l2, activation = relu), Dense (128, kernel_regularizer=l2, activation = relu), Dense (128, kernel_regularizer=l2, activation = relu), Dense (256, kernel_regularizer=l2, activation = relu),

Output Layer: Dense (1, activation=softplus)

Test Size: 15%

Best Accuracy = 75.79%

Best Combination = “softplus, adam, huber”

Worst Accuracy = 44.16%

Worst Combination = “softplus, sgd, huber”

Average Accuracy = 69.29%

Input Layer: Dense (64, kernel_regularizer=l2, activation = relu, input_shape = (30,))

Deep Layers: Dense (128, kernel_regularizer=l2, activation = relu), Dense (128, kernel_regularizer=l2, activation = relu), Dense (128, kernel_regularizer=l2, activation = relu),

Output Layer: Dense (1, activation=softplus)

Test Size: 15%

Best Accuracy = 98.78%

Best Combination = “softplus, adamax, huber”

Worst Accuracy = 44.65%

Worst Combination = “softplus, sgd, hinge”

Average Accuracy = 82.89%

Best Optimizers = rmsprop, adam, adamax, adadelta, adagrad, nadam

Worst Optimizers = SGD, ftrl

Best Activation Functions = softplus, sigmoid, selu, elu, exponential, tanh, softsign

Worst Activation Functions = relu

Best Combinations = selu, adamax, binary_crossentropy; %98.86

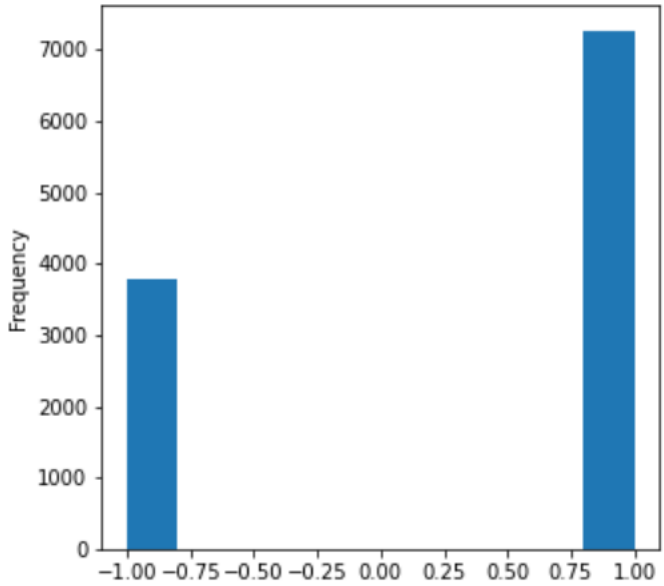
Worst Combinations = softplus, sgd, huber; %44.41

Average Accuracy = %81.42

IX. DATA SET FEATURES

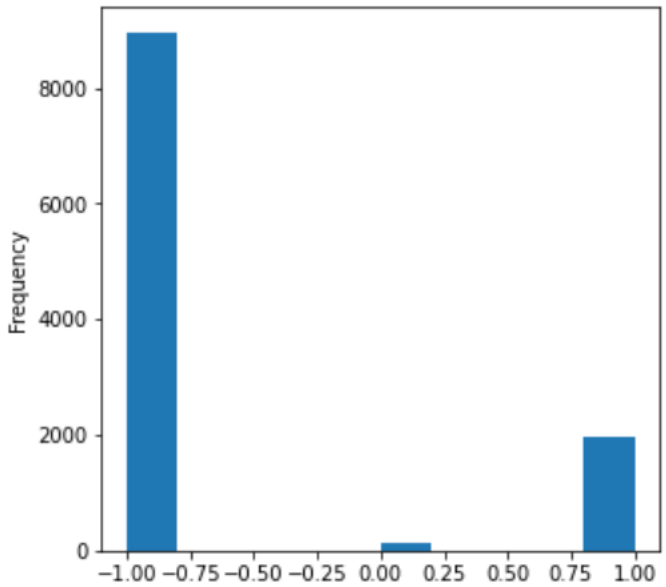
1. Using the IP Address

Rule: IF $\begin{cases} \text{If The Domain Part has an IP Address} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$



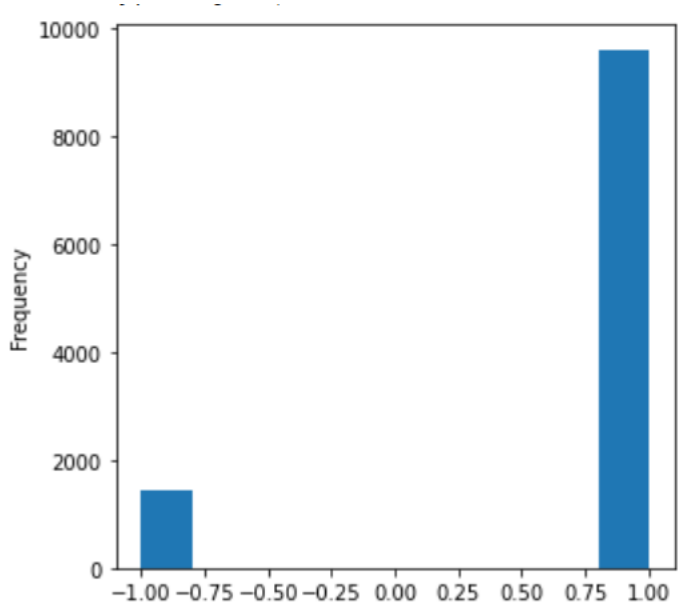
2. Long URL to Hide the Suspicious Part

Rule: IF $\begin{cases} URL\ length < 54 \rightarrow feature = \text{Legitimate} \\ else\ if\ URL\ length \geq 54\ and\ \leq 75 \rightarrow feature = Suspicious \\ otherwise \rightarrow feature = \text{Phishing} \end{cases}$



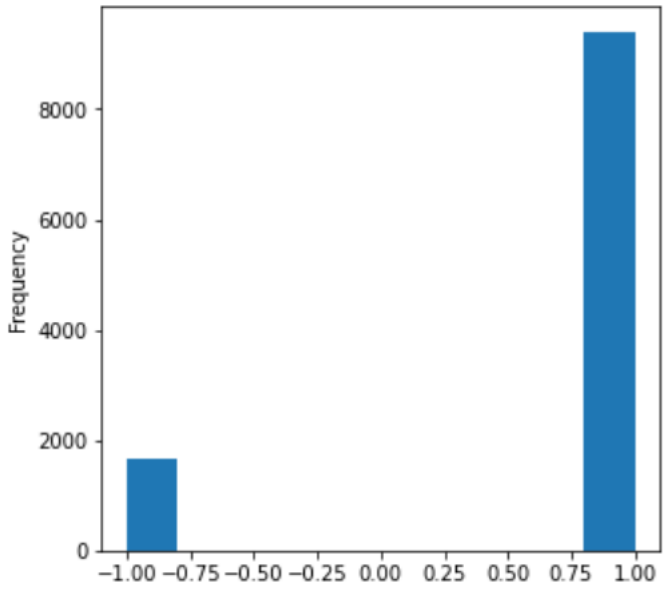
3. Using URL Shortening Services "TinyURL"

Rule: IF $\begin{cases} \text{TinyURL} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$



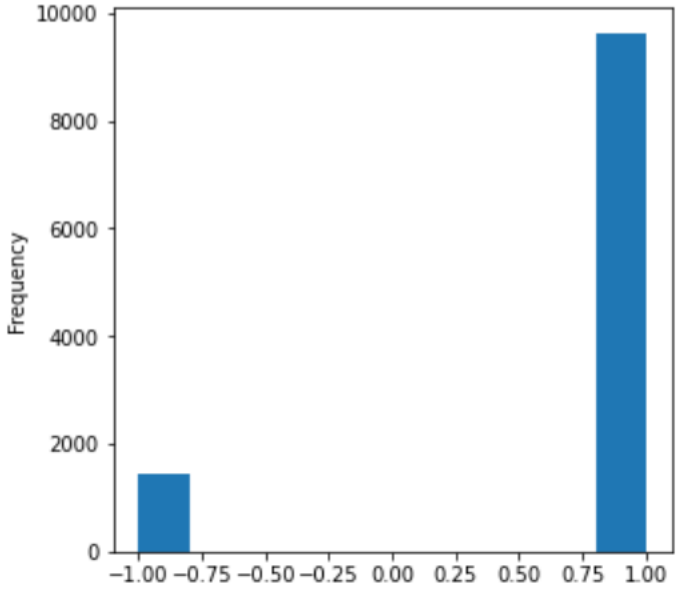
4. URL's having "@" Symbol

Rule: IF $\begin{cases} \text{Url Having @ Symbol} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$



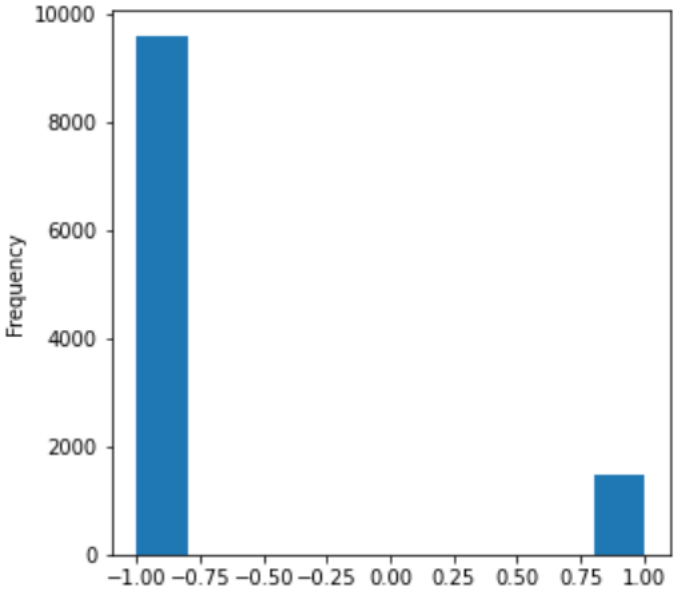
5. Redirecting using "//"

Rule: IF $\begin{cases} \text{ThePosition of the Last Occurrence of "//" in the URL} > 7 \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$



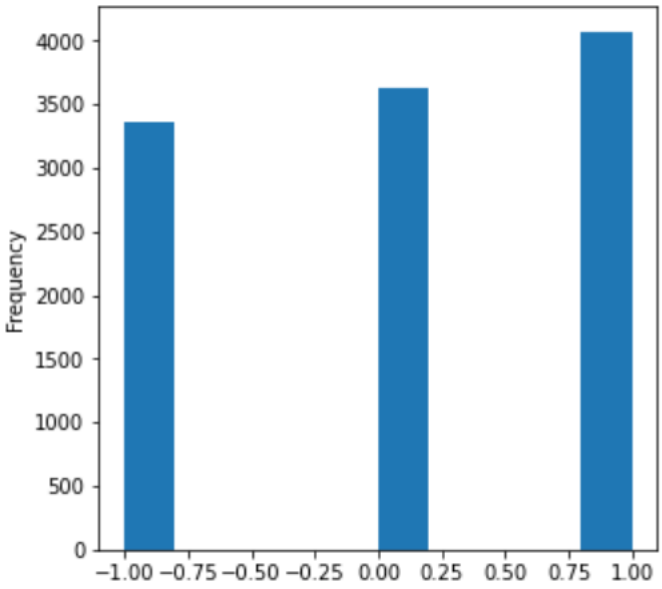
6. Adding Prefix or Suffix Separated by (-) to the Domain

Rule: IF $\begin{cases} \text{Domain Name Part Includes } (-) \text{ Symbol} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$



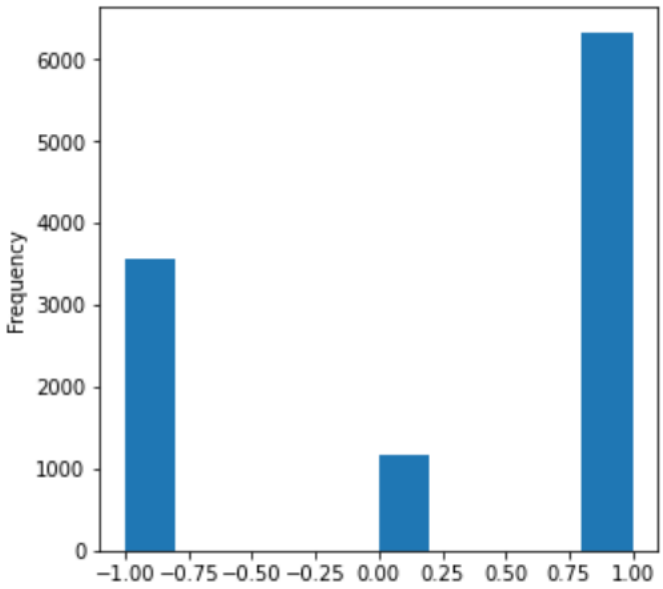
7. Sub Domain and Multi Sub Domains

Rule: IF $\begin{cases} \text{Dots In Domain Part} = 1 \rightarrow \text{Legitimate} \\ \text{Dots In Domain Part} = 2 \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases}$



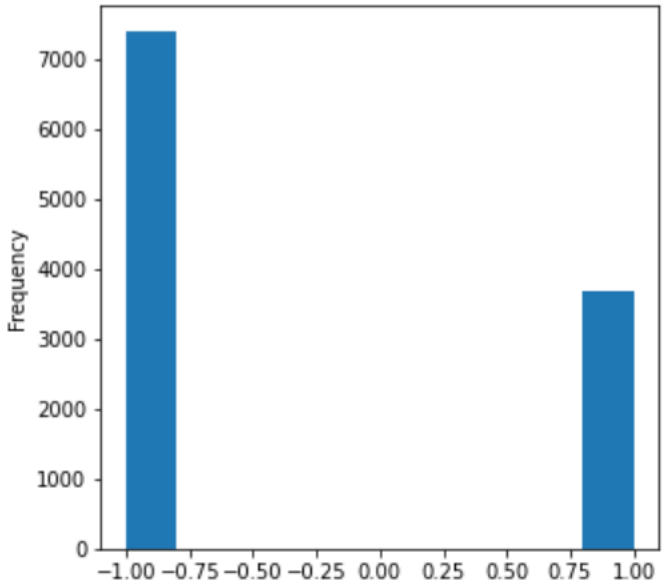
8. HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)

Rule: IF $\begin{cases} \text{Use https and Issuer Is Trusted and Age of Certificate} \geq 1 \text{ Years} \rightarrow \text{Legitimate} \\ \text{Using https and Issuer Is Not Trusted} \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases}$



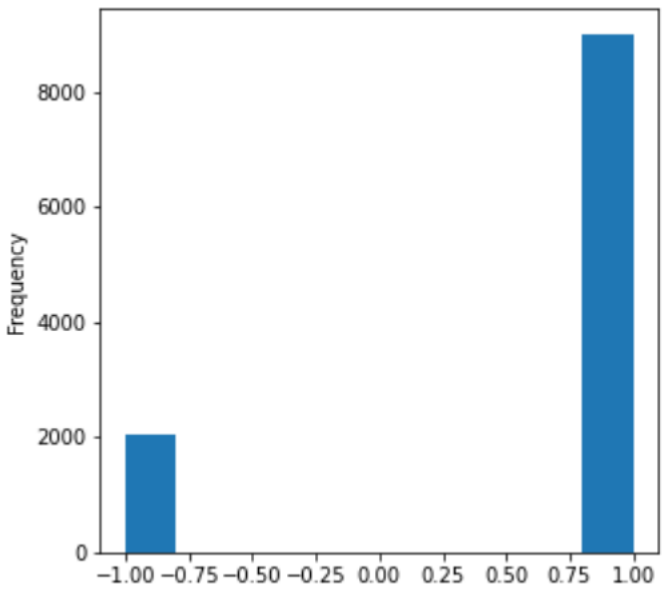
9. Domain Registration Length

Rule: IF $\begin{cases} \text{Domains Expires on} \leq 1 \text{ years} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$



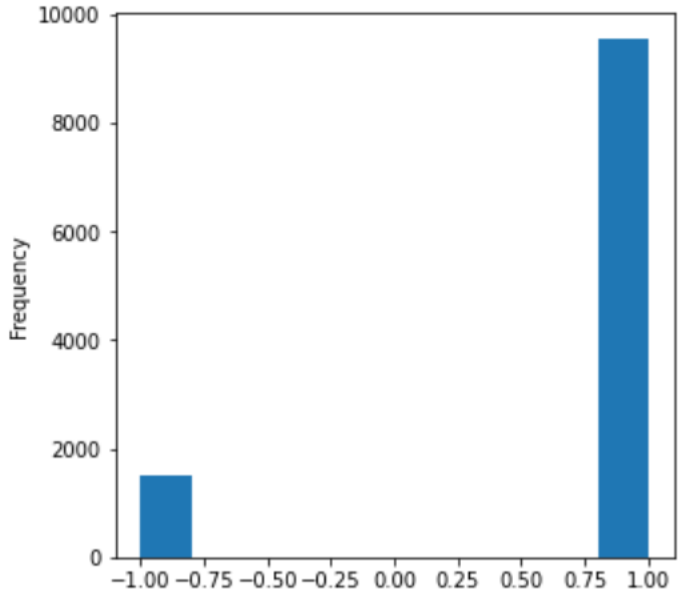
10. Favicon

Rule: IF $\begin{cases} \text{Favicon Loaded From External Domain} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$



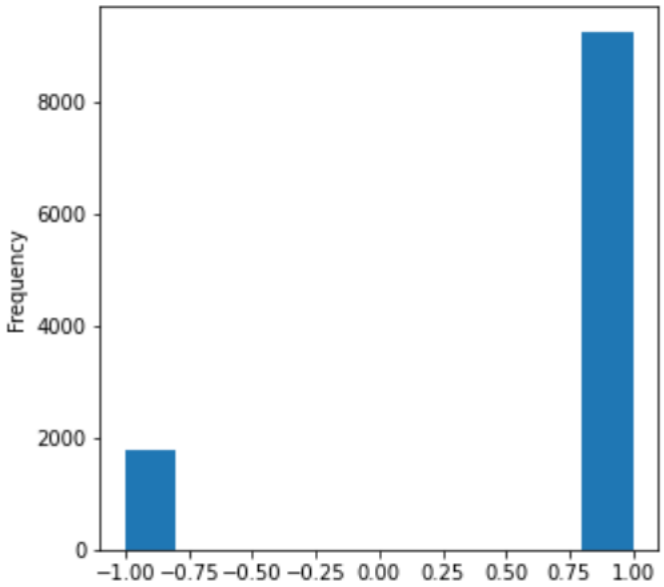
11. Using Non-Standard Port

Rule: IF $\begin{cases} \text{Port \# is of the Preffered Status} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$



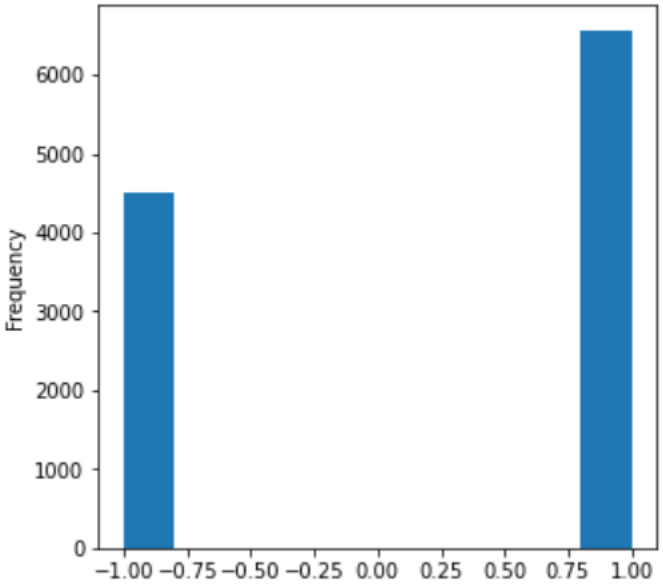
12. The Existence of "HTTPS" Token in the Domain Part of the URL

Rule: IF $\begin{cases} \text{Using HTTP Token in Domain Part of The URL} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$



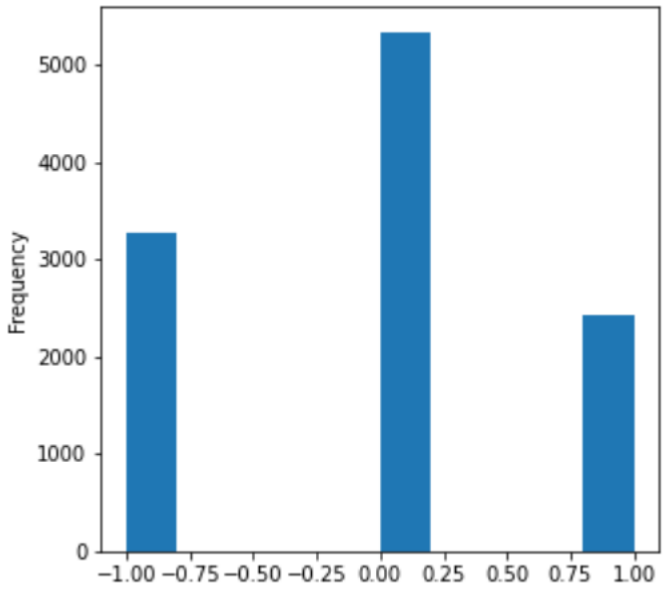
13. Request URL

Rule: IF $\begin{cases} \%\text{ of Request URL } < 22\% \rightarrow \text{Legitimate} \\ \%\text{of Request URL} \geq 22\% \text{ and } 61\% \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{feature} = \text{Phishing} \end{cases}$



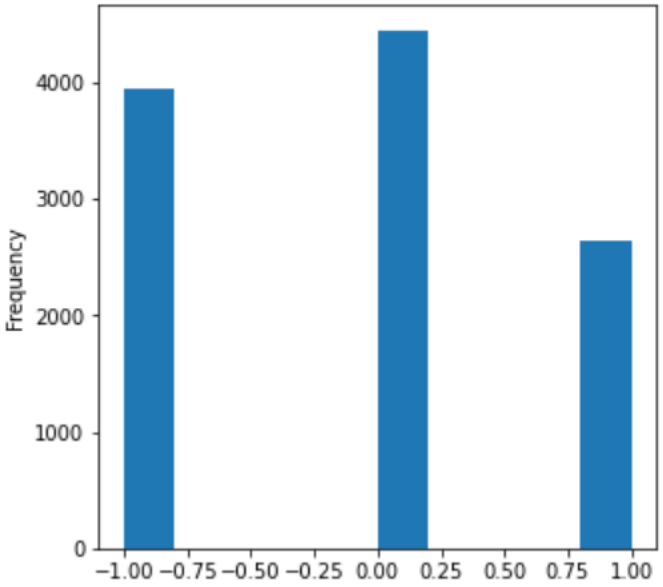
14. URL of Anchor

Rule: IF $\begin{cases} \%\text{ of URL Of Anchor } < 31\% \rightarrow \textit{Legitimate} \\ \%\text{ of URL Of Anchor } \geq 31\% \text{ And} \leq 67\% \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases}$



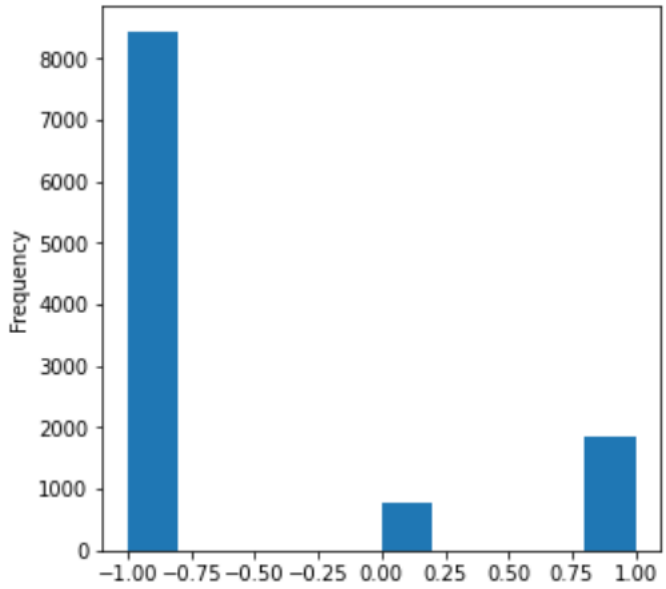
15. Links in <Meta>, <Script> and <Link> tags

$$\text{IF}\begin{cases} \%\text{ of Links in "} < \text{Meta} > \text{","} < \text{Script} > \text{" and "} < \text{Link>"} < 17\% \rightarrow \text{Legitimate} \\ \%\text{ of Links in } < \text{Meta} > \text{","} < \text{Script} > \text{" and "} < \text{Link>"} \geq 17\% \text{ And} \leq 81\% \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases}$$



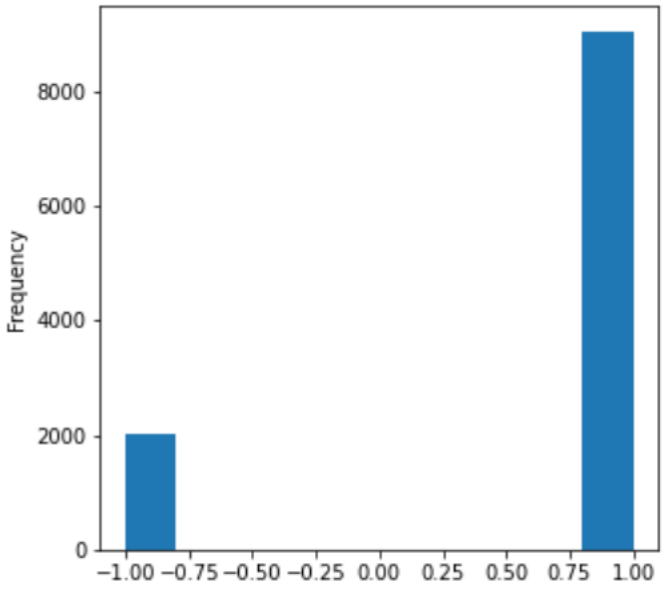
16. Server Form Handler (SFH)

$$\text{Rule: IF}\begin{cases} \text{SFH is "about: blank" Or Is Empty} \rightarrow \text{Phishing} \\ \text{SFH Refers To A Different Domain} \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$$



17. Submitting Information to Email

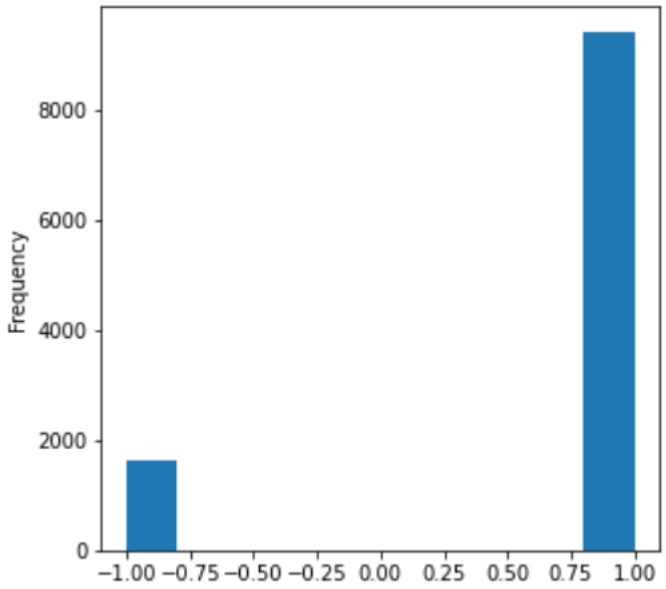
Rule: IF $\begin{cases} \text{Using "mail()" or "mailto:" Function to Submit User Information} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$



18. Abnormal URL

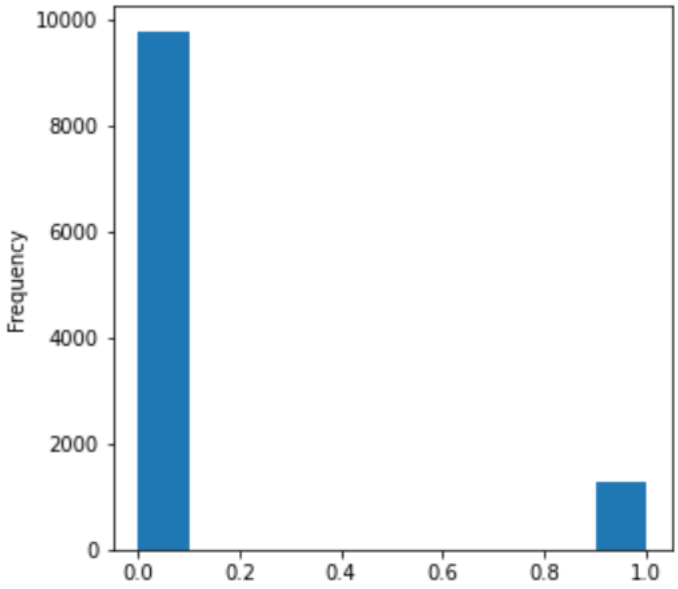
This feature can be extracted from WHOIS database. For a legitimate website, identity is typically part of its URL.

Rule: IF $\begin{cases} \text{The Host Name Is Not Included In URL} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$



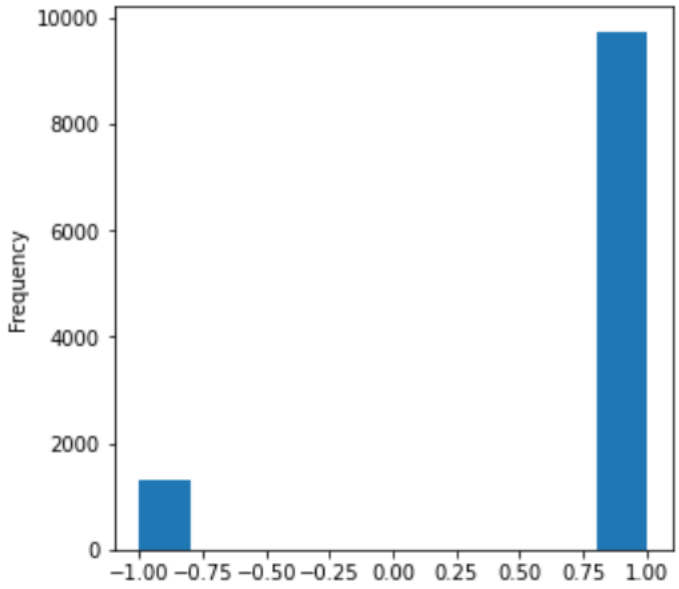
19. Website Forwarding

Rule: IF $\begin{cases} \text{ofRedirect Page} \leq 1 \rightarrow \text{Legitimate} \\ \text{of Redirect Page} \geq 2 \text{ And} < 4 \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases}$



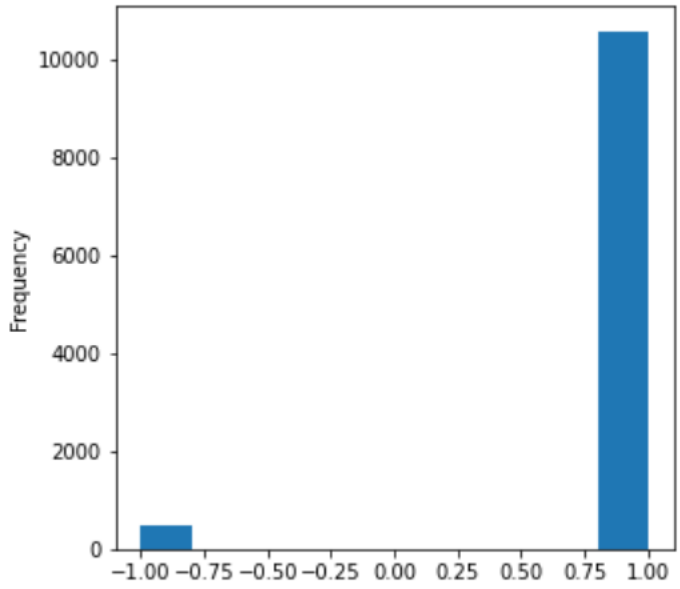
20. Status Bar Customization

Rule: IF $\begin{cases} \text{onMouseOver Changes Status Bar} \rightarrow \text{Phishing} \\ \text{It Does't Change Status Bar} \rightarrow \text{Legitimate} \end{cases}$



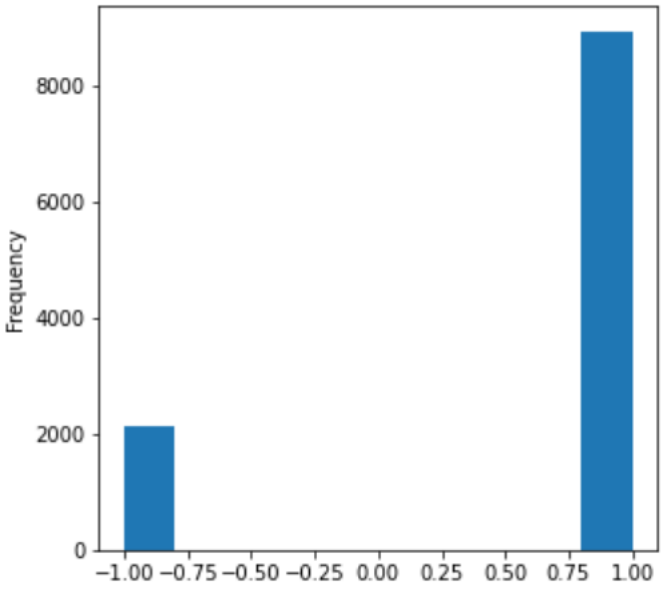
21. Disabling Right Click

Rule: IF $\begin{cases} \text{Right Click Disabled} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$



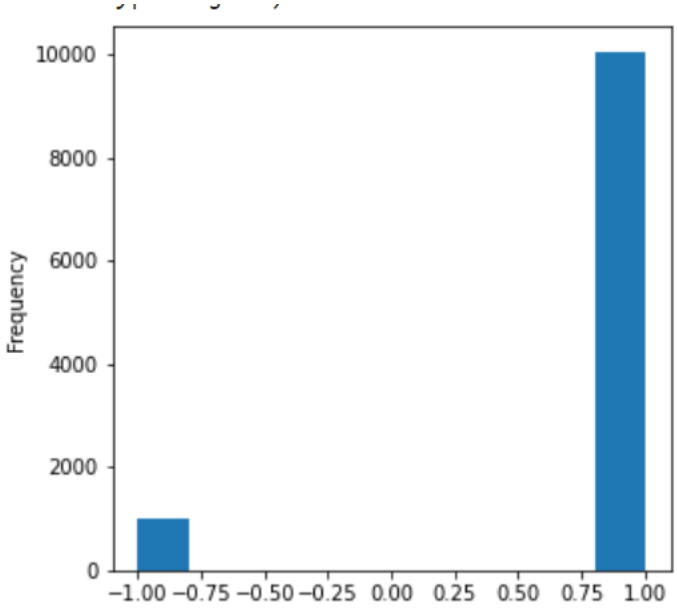
22. Using Pop-up Window

Rule: IF $\begin{cases} \text{Popoup Window Contains Text Fields} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$



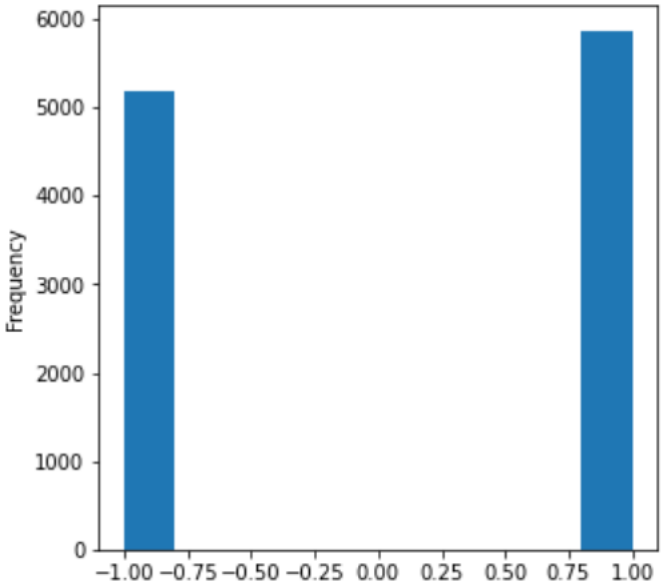
23. IFrame Redirection

Rule: IF $\begin{cases} \text{Using iframe} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$



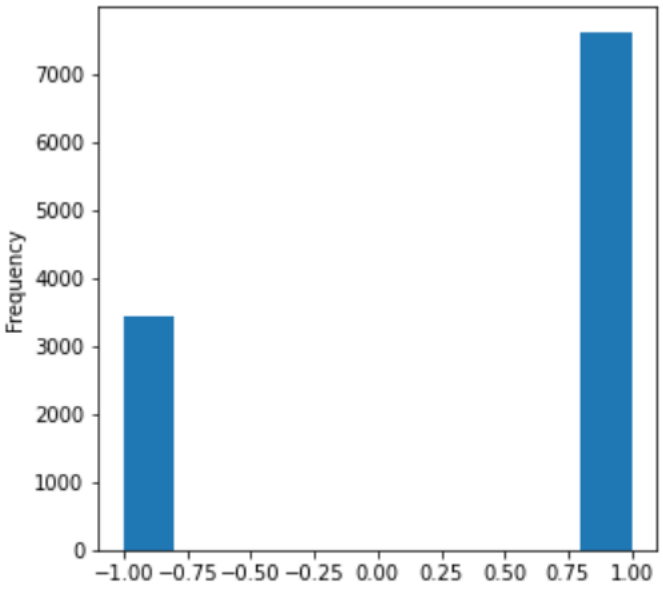
24. Age of Domain

Rule: IF $\begin{cases} \text{Age Of Domain} \geq 6 \text{ months} \rightarrow \text{Legitimate} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases}$



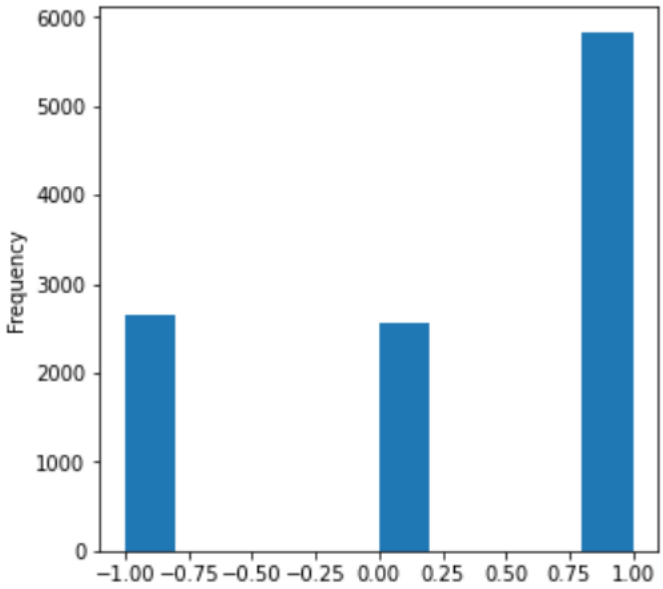
25. DNS Record

Rule: IF $\begin{cases} \text{no DNS Record For The Domain} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$



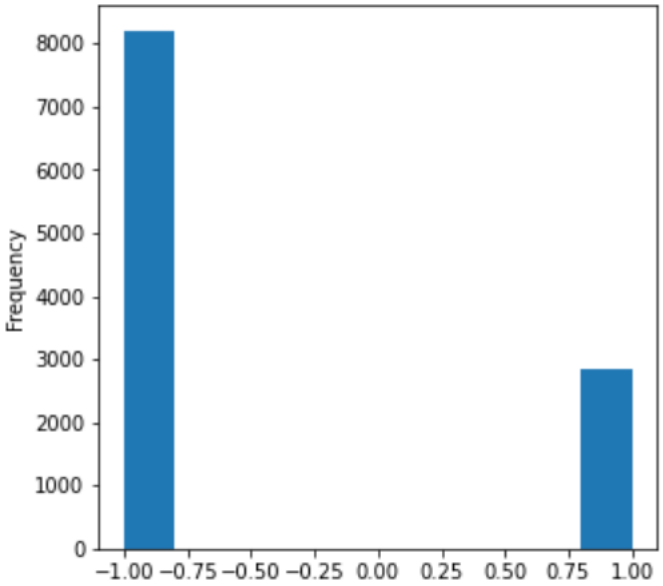
26. Website Traffic

Rule: IF $\begin{cases} \text{Website Rank} < 100{,}000 \rightarrow \text{Legitimate} \\ \text{Website Rank} > 100{,}000 \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phish} \end{cases}$



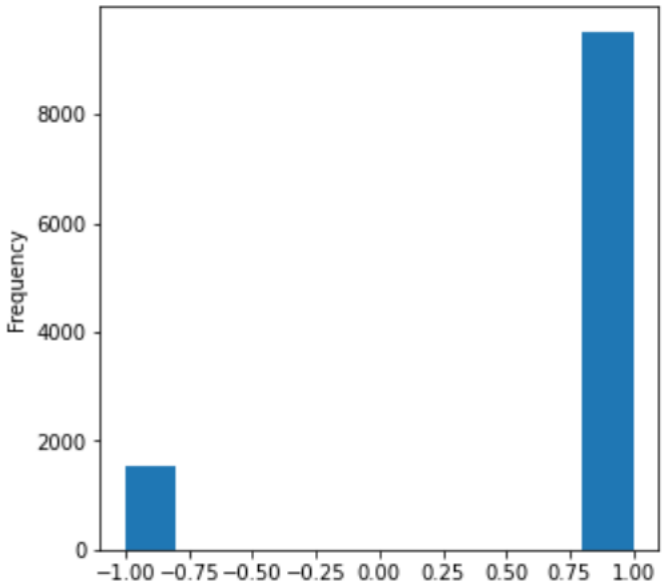
27. PageRank

Rule: IF $\begin{cases} \text{PageRank} < 0.2 \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$



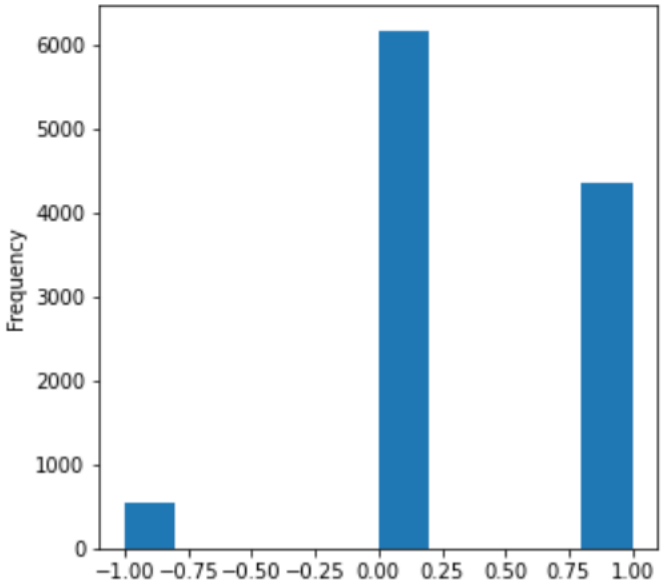
28. Google Index

Rule: IF $\begin{cases} \text{Webpage Indexed by Google} \rightarrow \text{Legitimate} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases}$



29. Number of Links Pointing to Page

Rule: IF $\begin{cases} \text{Of Link Pointing to The Webpage} = 0 \rightarrow \text{Phishing} \\ \text{Of Link Pointing to The Webpage} > 0 \text{ and} \leq 2 \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$



30. Statistical-Reports Based Feature

Rule: IF $\begin{cases} \text{Host Belongs to Top Phishing IPs or Top Phishing Domains} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$