BookMyShow Privacy Security Analysis

Group 2

01

Data Selection
& Mission Statement

02

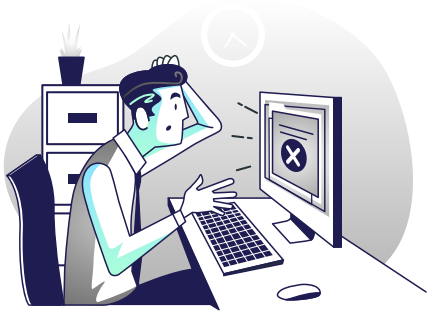
Data Preprocessing
& Exploration Analysis

03

Model Selection

Data Selection

The data set we selected is the collection of URL ads (11k sample) that included 32 features could be used as classifiers.



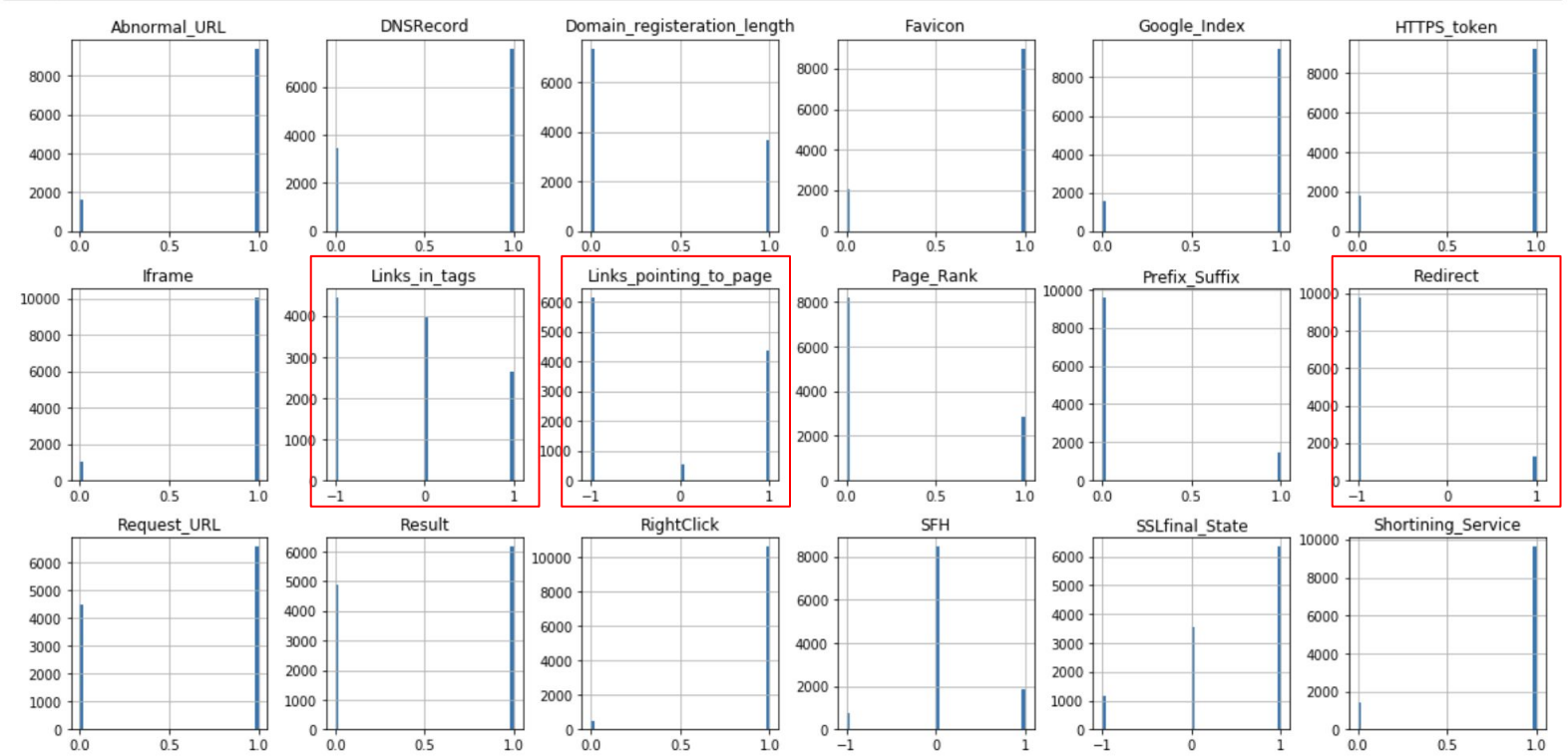
Mission Statement

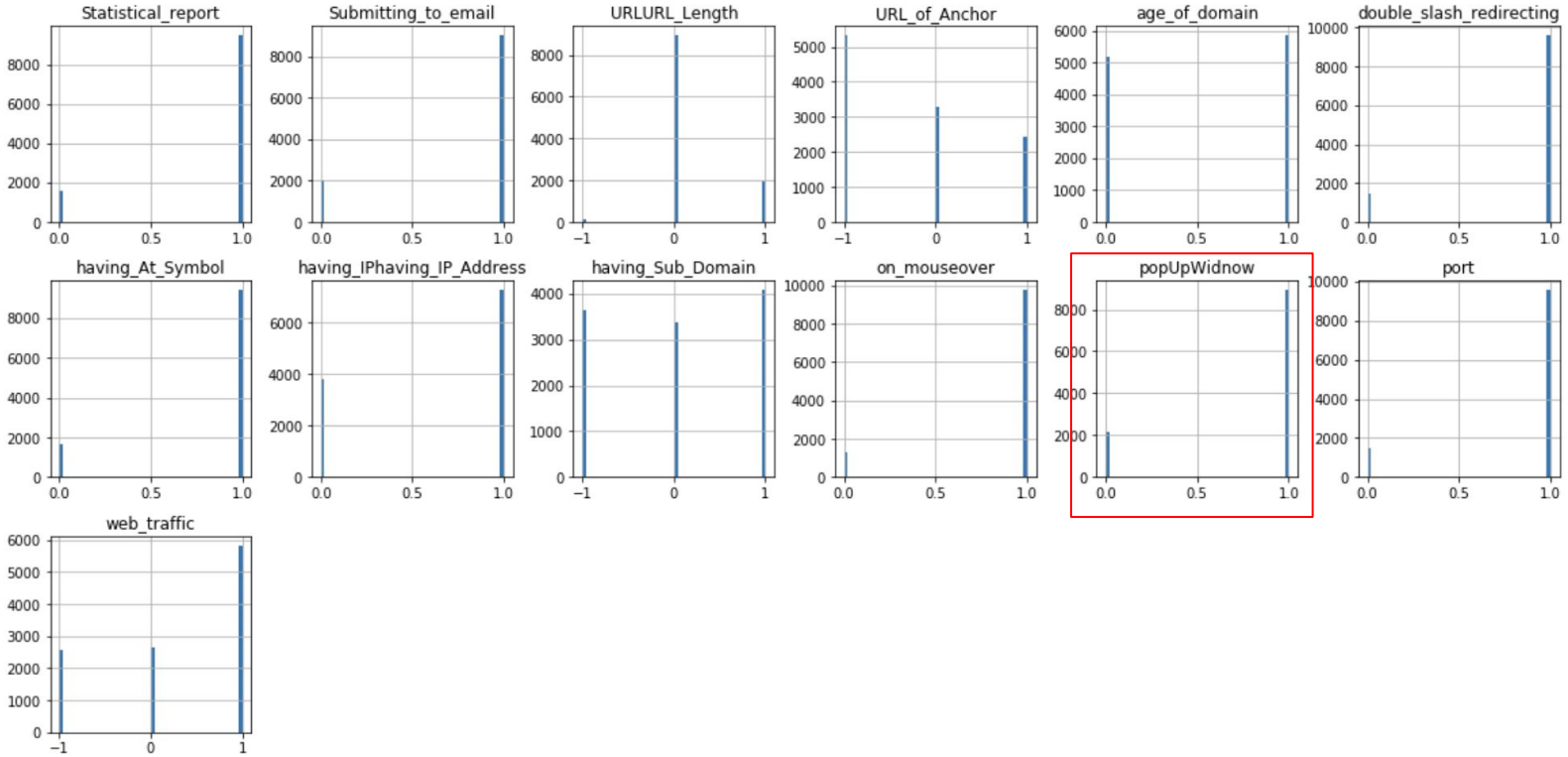
For privacy security, we want to classify whether a particular URL ads is prone to phishing or not

Data Preprocessing

- 32 features are categorical variables
  - 0 = Phishing
  - 1 = Legitimate
  - -1 = Suspicious
- Standardization is not needed
- After removing missing values, we left with 11055 obs., 32 features

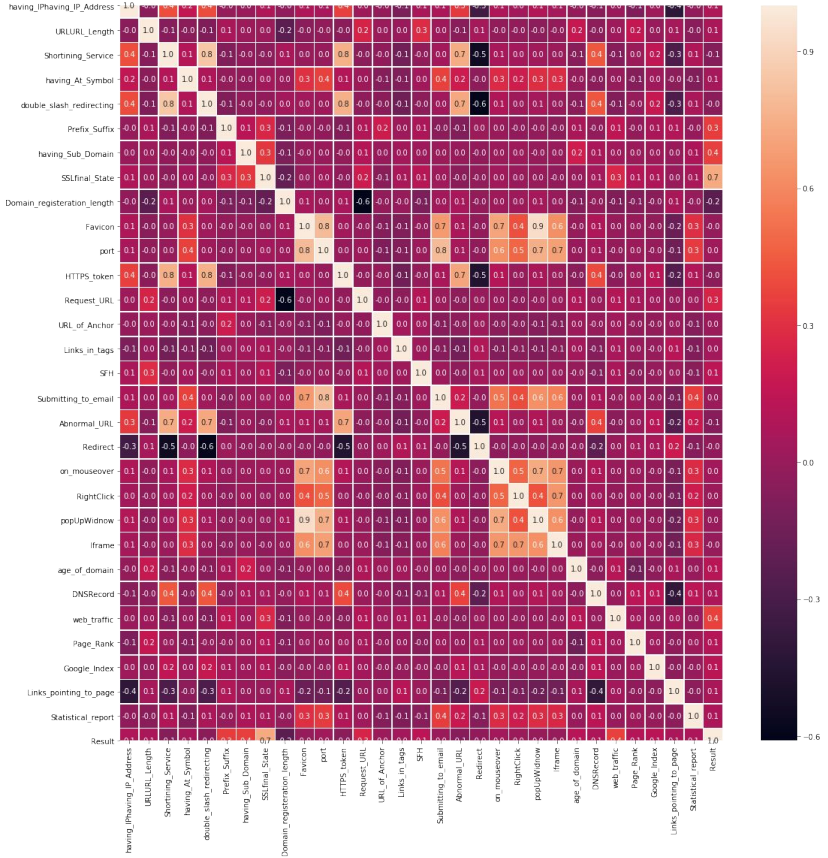
Data Exploration - Histogram




Statistical_report
Submitting_to_email
URLURL_Length
URL_of_Anchor
age_of_domain
double_slash_redirecting
having_At_Symbol
having_IPhaving_IP_Address
having_Sub_Domain
on_mouseover
popUpWidnow
port
web_traffic

Correlation - Heatmap

- Removed multicollinearity features with 0.8 threshold (removed features: 'double_slash_redirecting', 'port', 'popUpWidnow')
- After removing, we left with 29 variables



Model Selection

70% training, 30% Testing, seed(123)

- **Binary Classification** that classifies whether the URL sample is a phishing site or not.
    - Logistic Regression
    - KNN Classifier
    - Decision Tree
    - Random Forest

Logistic Regression

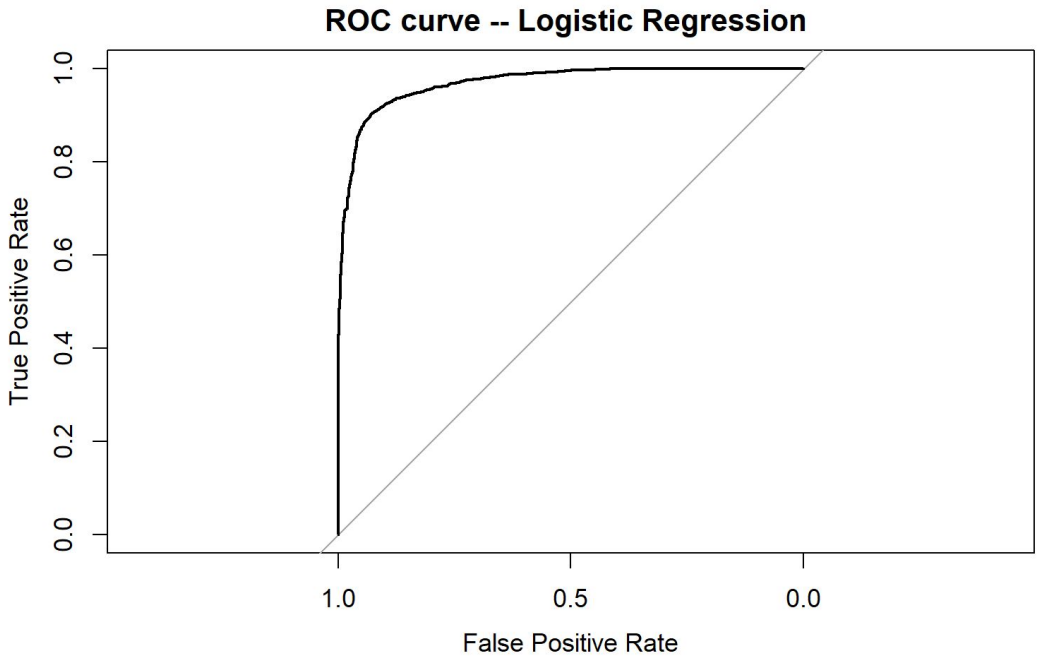
RStudio: Notebook Output

```
Call:
glm(formula = Result ~ ., family = binomial(), data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.86665  -0.22529   0.00002   0.29067   3.09787

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -4.86870    0.33962 -14.336  < 2e-16 ***
having_IPhaving_IP_Address  1.51721    0.13116  11.568  < 2e-16 ***
URLURL_Length               0.15898    0.13382   1.188  0.23483
Shortining_Service         -1.12039    0.26306  -4.259 2.05e-05 ***
having_At_Symbol            0.66349    0.16045   4.135 3.55e-05 ***
Prefix_Suffix              18.91094  270.46794   0.070  0.94426
having_Sub_Domain           0.88115    0.06041  14.585  < 2e-16 ***
SSLfinal_State              3.85280    0.09958  38.689  < 2e-16 ***
Domain_registeration_length -0.51745    0.11841  -4.370 1.24e-05 ***
Favicon                    -0.47404    0.19399  -2.444  0.01454 *
HTTPS_token                -0.50260    0.19956  -2.519  0.01179 *
Request_URL                 0.49876    0.11417   4.368 1.25e-05 ***
URL_of_Anchor               0.08594    0.06227   1.380  0.16754
Links_in_tags              -0.32102    0.06098  -5.264 1.41e-07 ***
SFH                         0.32347    0.10633   3.042  0.00235 **
Submitting_to_email         0.01628    0.17009   0.096  0.92373
Abnormal_URL               -0.64745    0.23731  -2.728  0.00637 **
Redirect                   -0.44937    0.09238  -4.864 1.15e-06 ***
on_mouseover                0.25260    0.25050   1.008  0.31327
RightClick                  0.19999    0.32185   0.621  0.53436
Iframe                      0.03091    0.29428   0.105  0.91633
age_of_domain               0.01535    0.09715   0.158  0.87444
DNSRecord                   1.62696    0.13380  12.160  < 2e-16 ***
web_traffic                 1.18429    0.05802  20.411  < 2e-16 ***
Page_Rank                   0.59676    0.11308   5.277 1.31e-07 ***
Google_Index                1.00622    0.13160   7.646 2.08e-14 ***
Links_pointing_to_page      0.78254    0.06285  12.451  < 2e-16 ***
Statistical_report          0.39392    0.16311   2.415  0.01573 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10628.0  on 7738  degrees of freedom
Residual deviance:  3379.9  on 7711  degrees of freedom
AIC: 3435.9
```

- **X variables**: 29 variables in total
  - Narrow down to 18 variables for final state
- **Y variables**: “Result” variable with two classes
  - 0 being Phishing
  - 1 being Legitimate
- **Classification Accuracy**:
  - 91.45 on training
  - 90.53 on validation
- **F-score**: 0.92

Logistic Regression

Area under the curve: 0.9697



KNN

```
k-Nearest Neighbors

11055 samples
   27 predictor
    2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 9950, 9949, 9949, 9950, 9949, 9950, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   1  0.9692429  0.9375609
   2  0.9594732  0.9177026
   3  0.9547701  0.9081144
   4  0.9497944  0.8980715
   5  0.9479851  0.8944383
   6  0.9458145  0.8900164
   7  0.9443670  0.8870684
   8  0.9412012  0.8806803
   9  0.9405673  0.8794050
  10  0.9411109  0.8805174

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 1.
```

- **X variables**: all 28 variables.
- **Y variables**: Result
  - 0 being Phishing
  - 1 being Legitimate
- **Classification Accuracy**:
  - Best accuracy when k=1
  - Since our data set is all binary, there is no problem of overfitting
  - However, it also indicates that KNN classifier is not our best model, since there is not much to explain.

Classification Tree

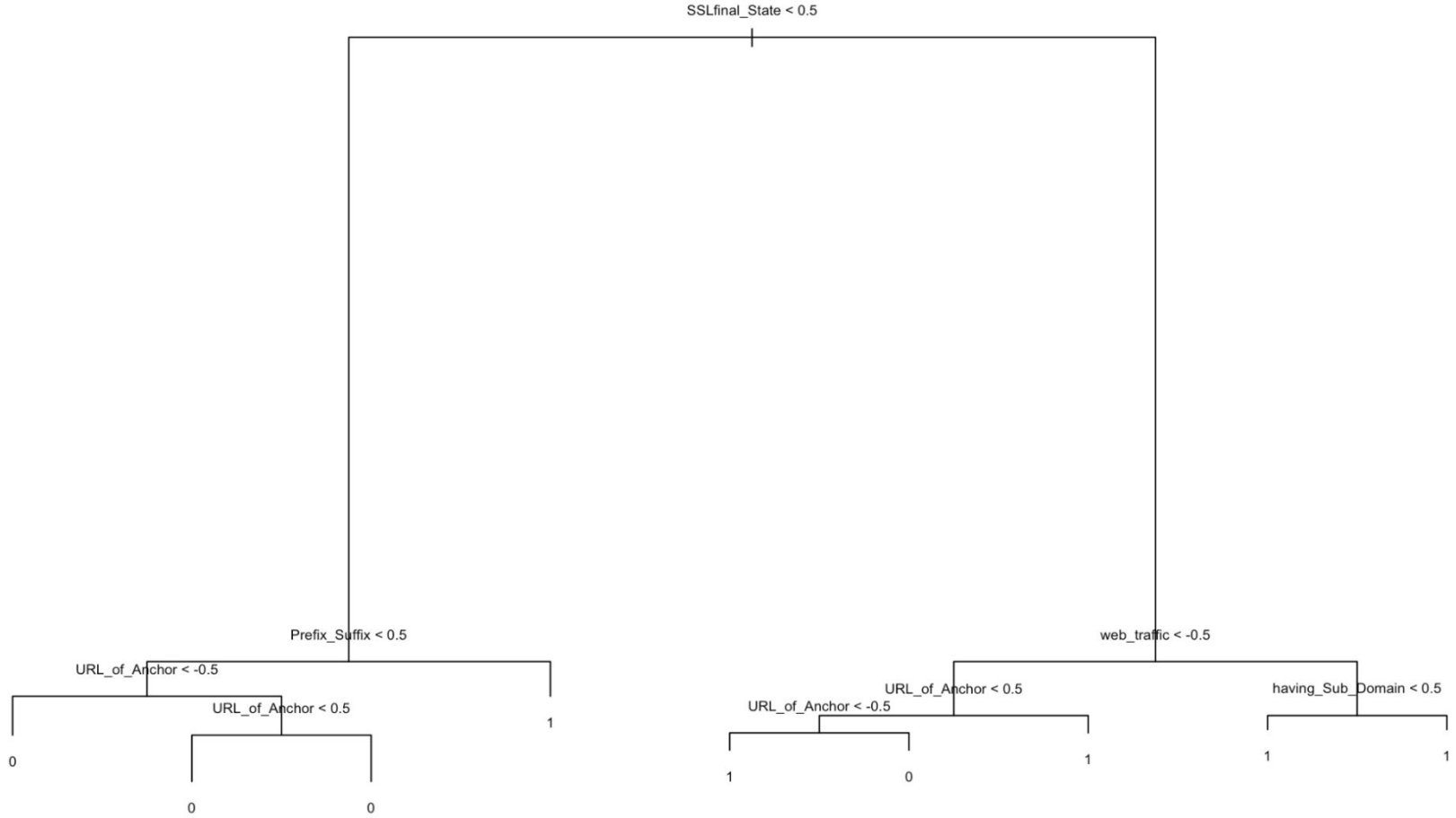
```
Classification tree:
tree(formula = Result ~ ., data = train)
Variables actually used in tree construction:
[1] "SSLfinal_State"    "Prefix_Suffix"     "URL_of_Anchor"     "web_traffic"
[5] "having_Sub_Domain"
Number of terminal nodes:  9
Residual mean deviance:  0.4494 = 3474 / 7730
Misclassification error rate: 0.09136 = 707 / 7739
```

Despite there being 28 predictors in the dataset, **only 5 were used in splits**.
These were:

- "SSLfinal_State"
- "Prefix_Suffix"
- "URL_of_Anchor"
- "web_traffic"
- "having_Sub_Domain"

Classification Accuracy: 90.44
Error Rate: 0.0955

```
          test_actual
test_pred    0    1
        0 1276  124
        1  193 1723
[1] 0.9044029
```


SSLfinal_State < 0.5
Prefix_Suffix < 0.5
URL_of_Anchor < -0.5
URL_of_Anchor < 0.5
0
0
0
1
web_traffic < -0.5
URL_of_Anchor < 0.5
URL_of_Anchor < -0.5
having_Sub_Domain < 0.5
1
0
1
1
1

Random Forest

- **Number of trees:** 500
- **No. of variables tried at each split:** 5

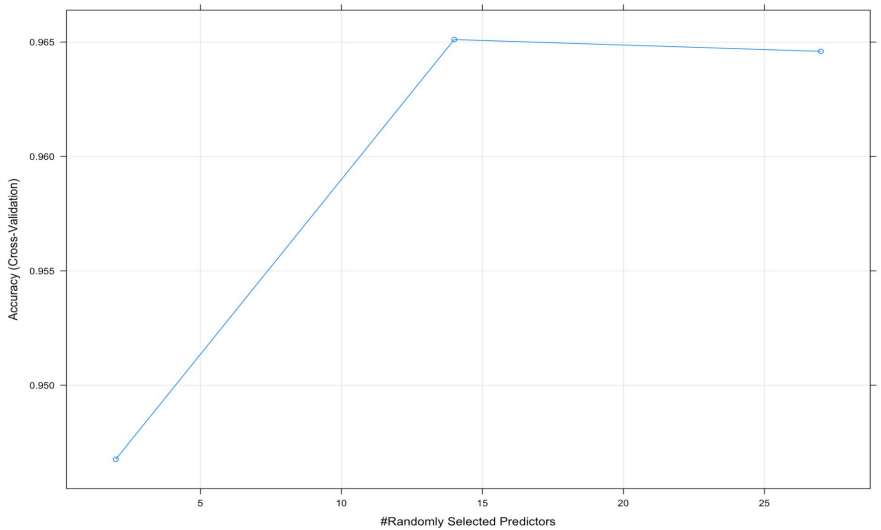
- **Classification Accuracy**:
  - 98.04 on training
  - 96.53 on validation

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1399   45
         1   70 1802

               Accuracy : 0.9653
                 95% CI : (0.9585, 0.9713)
    No Information Rate : 0.557
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9296

 Mcnemar's Test P-Value : 0.02522

            Sensitivity : 0.9756
            Specificity : 0.9523
         Pos Pred Value : 0.9626
         Neg Pred Value : 0.9688
             Prevalence : 0.5570
         Detection Rate : 0.5434
   Detection Prevalence : 0.5645
      Balanced Accuracy : 0.9640

       'Positive' Class : 1
```

Random Forest Cont'd

10-fold cross-validation

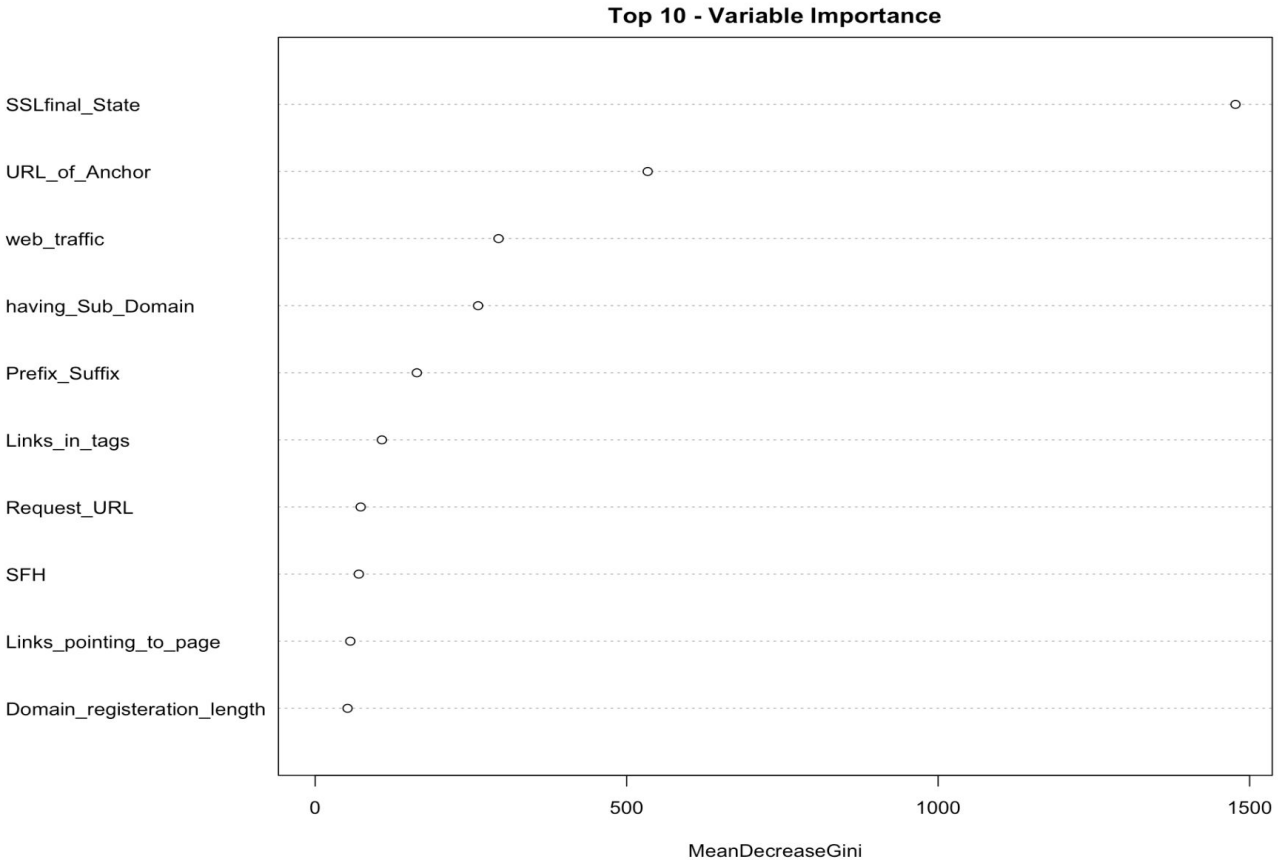
```
Random Forest

7739 samples
  27 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 6965, 6965, 6965, 6965, 6965, 6965, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.9467633  0.8918357
  14    0.9651109  0.9292071
  27    0.9645941  0.9281990

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 14.
```




```
             predictions
test.target    0    1
          0 1410   59
          1   36 1811
```

- **14 variables**
- **Accuracy on validation:** 97.13

Summary: RF As Our Final Model



Summary: Interpretation of Top 5 Important Variables

Importance (Decreasing Order)	Variable Name	Meaning
1	SSLfinal_State (Secure Sockets Layer)	Standard technology for keeping an internet connection secure and safeguarding any sensitive data that is being sent between two systems
2	URL_of_Anchor (The <a> HTML element)	Creates a hyperlink to web pages, or anything else a URL can address; If <a> tags ≠ the website domain names -> “Suspicious”
3	Web_traffic (Popularity of the website)	No traffic or is not recognized by the Alexa database -> “Phishing”

developer.mozilla.org/en-US/docs/Web/HTML/Element/a#:~:text=The%20HTML%20ele...

Top 5 Important Variables Cont'd

Importance (Decreasing Order)	Variable Name	Meaning
4	having_Sub_Domain (# of dots in domain name)	Count the remaining dots after removing "Country Code Top Level Domains". If # dots >1, URL = "Suspicious" If # dots > 2, URL = "Phishing"
5	Prefix_Suffix (Suffix Separated by (-) to the Domain)	The dash symbol is rarely used in legitimate URLs. Phishers tend to add prefixes or suffixes separated by (-) to the domain name so that users feel that they are dealing with a legitimate web-page

Thank you!

Questions or Comments?