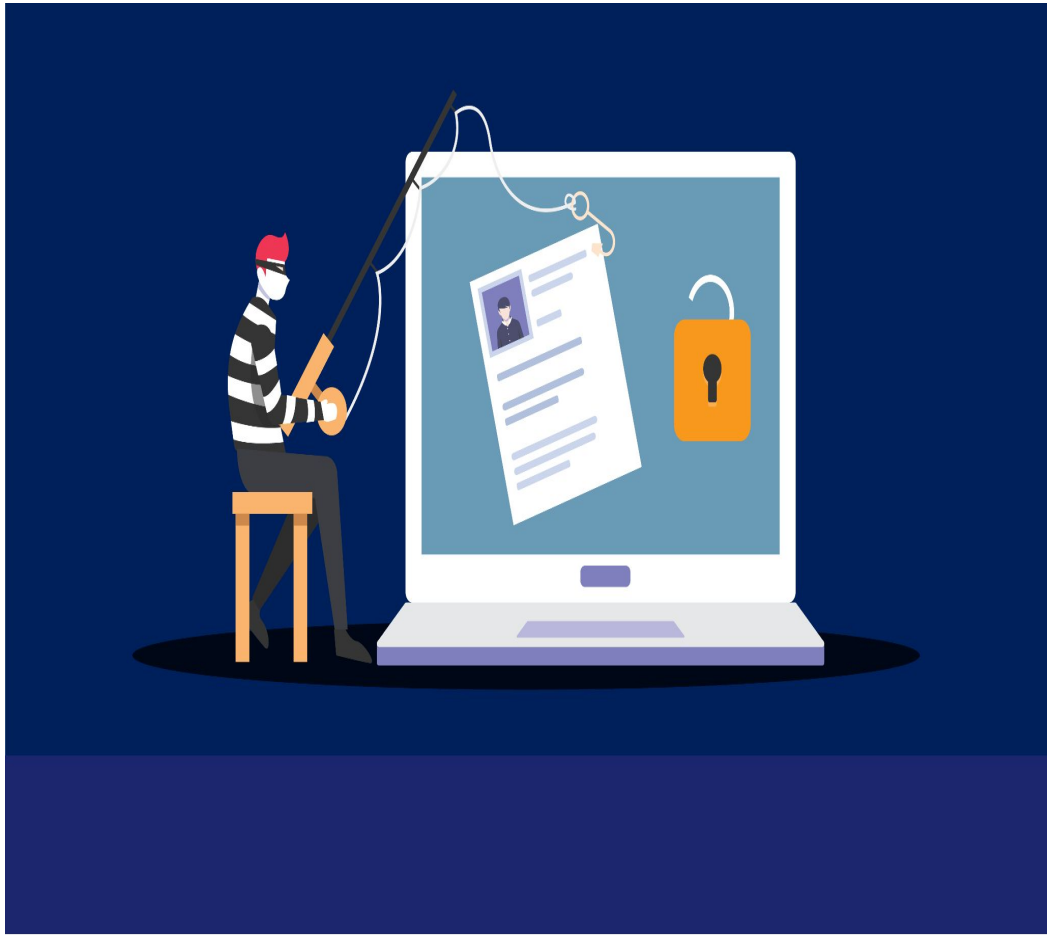
# Detection of Phishing URL


## Predictive Modeling
## Second Project Presentation- Model Building

**Group 14**
Mihret Kemal
Tagore Kosireddy
Saket Pawar

# Goal of the Project



- To classify URLs as either legitimate or phishing to help prevent phishing attacks

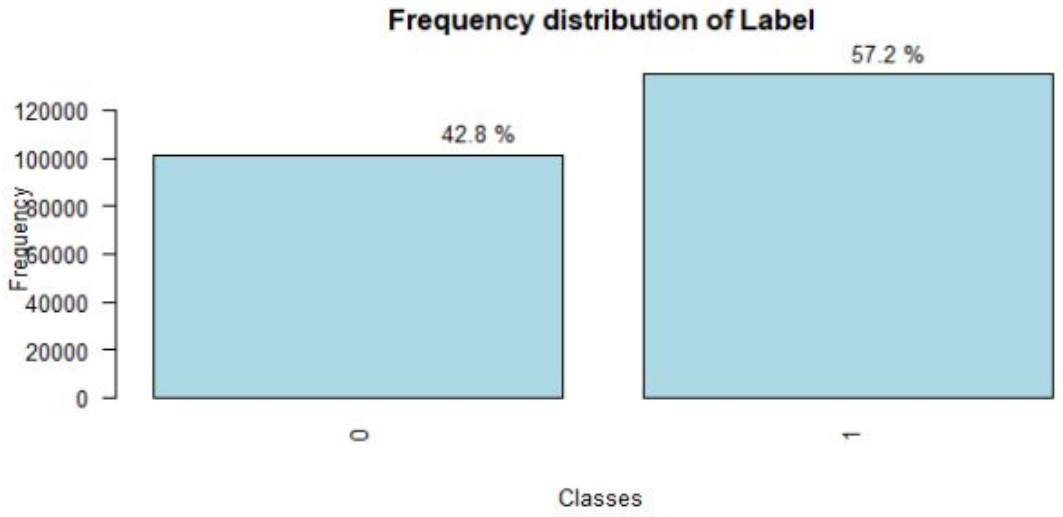
  why?

- According to Forbes, **over 500 million** phishing attacks were reported in 2022

# Dataset Description

- PhiUSIIL Phishing URL dataset from UCI machine learning repository is used.
- Sample size of the dataset is 235,795
- The dataset contains 56 columns totally (4 columns are extra information about url, namely : filename, URL, Domain, Title)

- Got 51 predictors - 33 continuous and 18 categorical
- The response value is called has 2 classes
- Label 1 - Legitimate URL
- Label 0 - Phishing URL



# Data Preprocessing

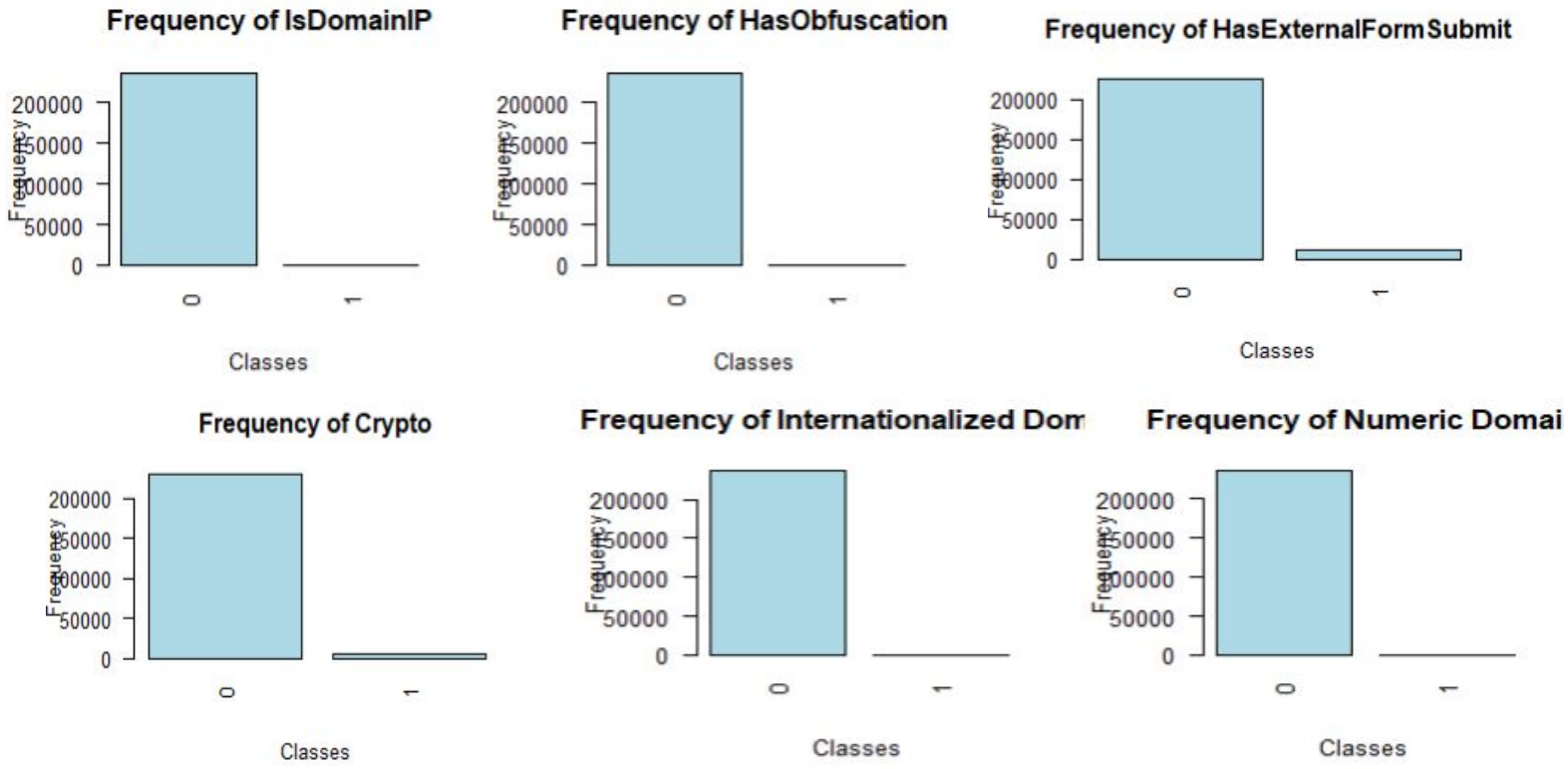
- The dataset contains no missing value(no imputation)

```
> total_missing <- sum(is.na(data))
> # Print the total number of missing values
> total_missing
[1] 0
>
```

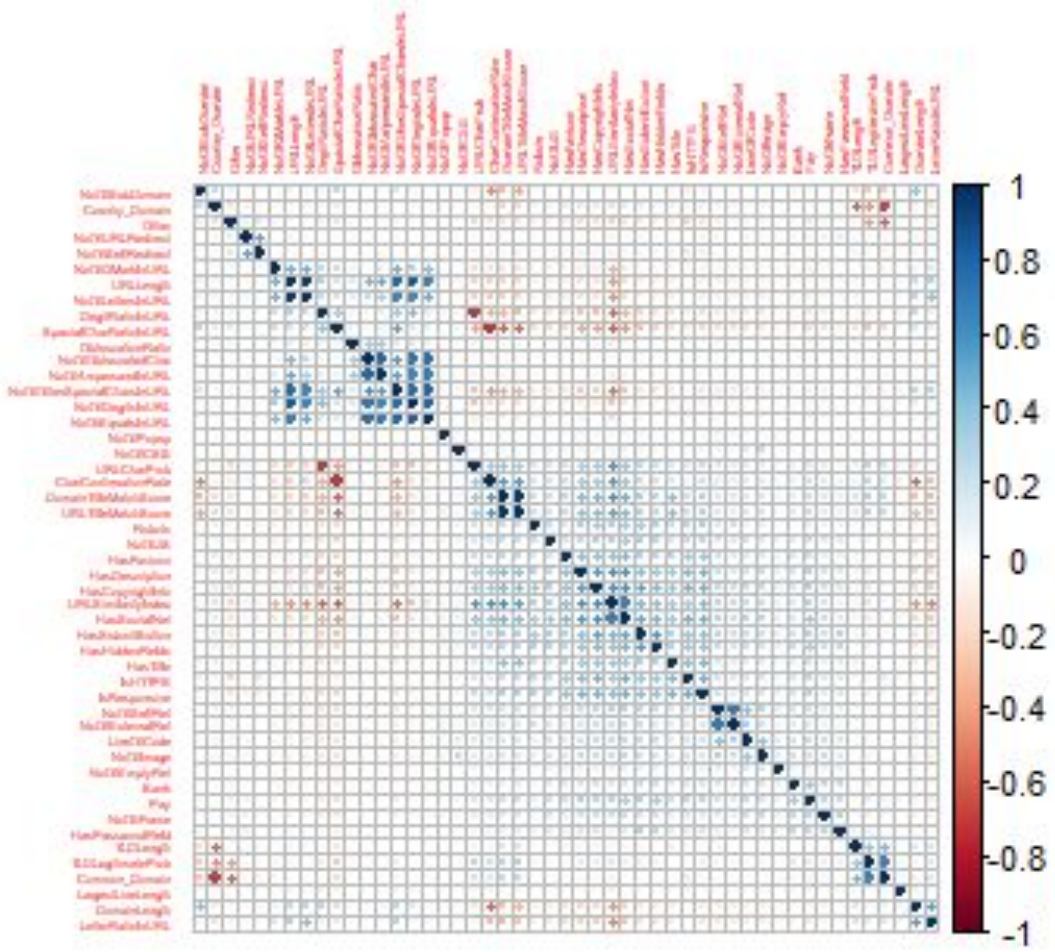
- Added 5 dummy variables from TLD predictor namely: Internationalized_domain, Country_domain, Common_domain, Number_domain

# Removal of Near-zero variance predictors

- Got 6 Near-zero variance predictors, so we deleted them



# Any Correlation?



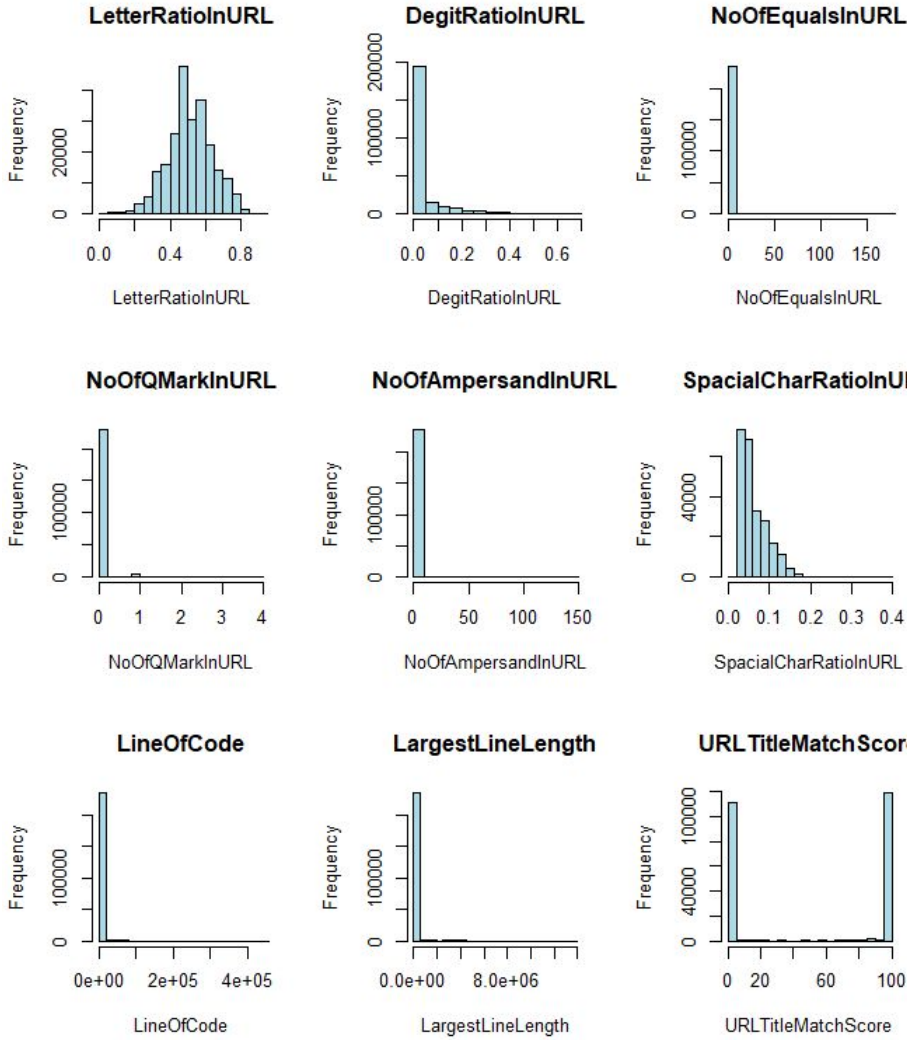
# Removal of highly correlated predictors

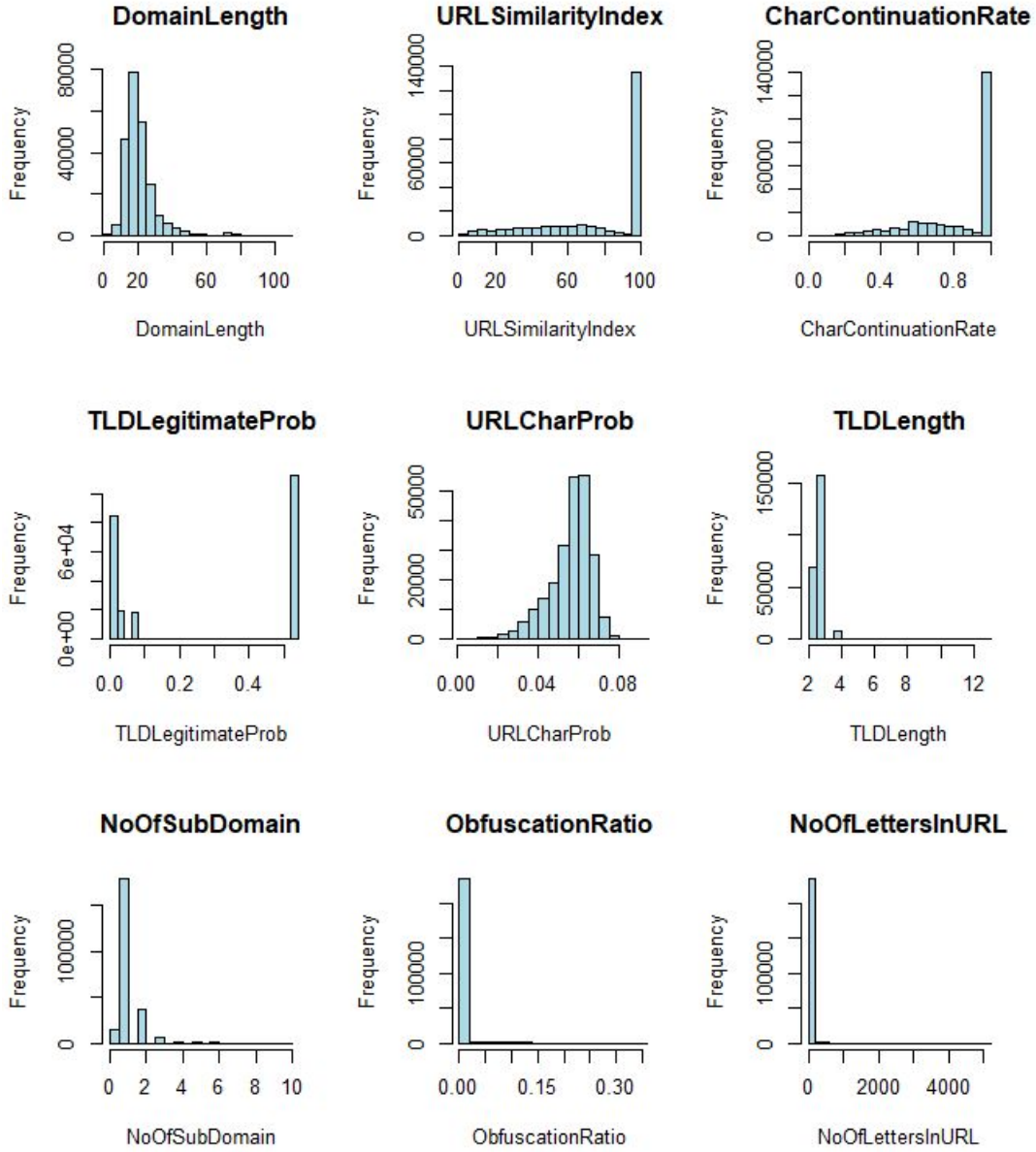
- Checked for highly correlated predictors (cutoff = 0.75), and found 5

```
> colnames(all_num_data)[highCorr]
[1] "DomainTitleMatchScore"      "NoOfOtherSpecialCharsInURL" "URLLength"
[4] "NoOfDegitsInURL"            "NoOfObfuscatedChar"
>
```

- So We deleted this 5 columns

# Distributions

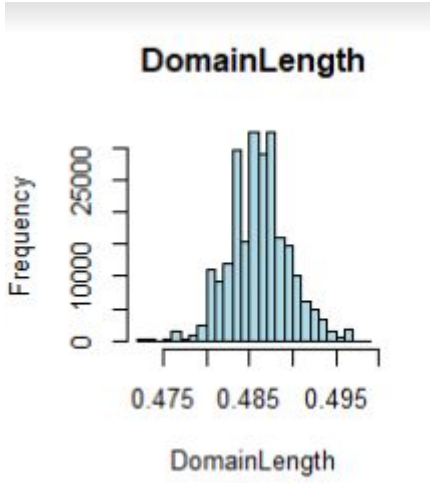


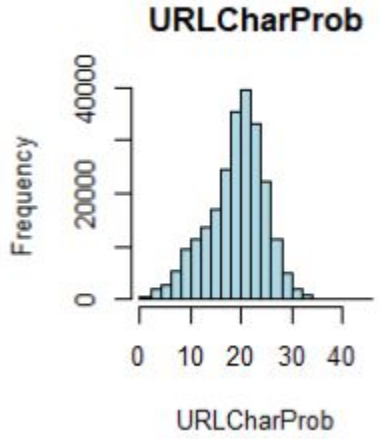



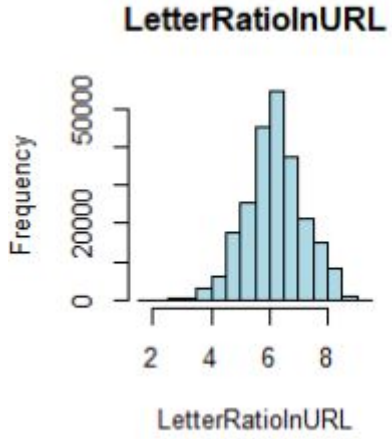


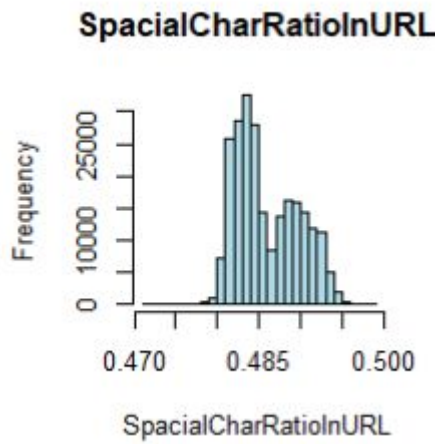

Michigan Tech

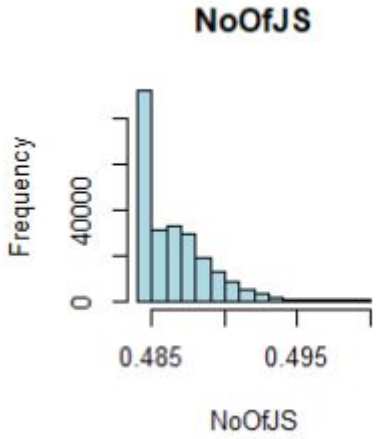
# Applied Box-Cox Transformation



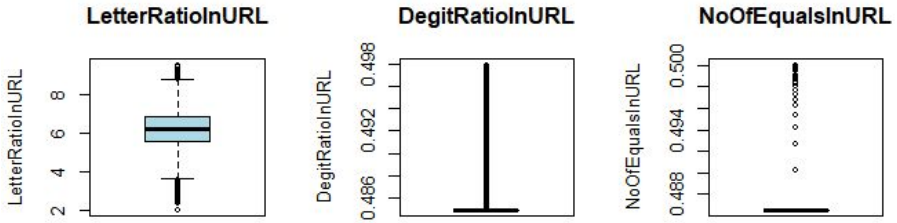



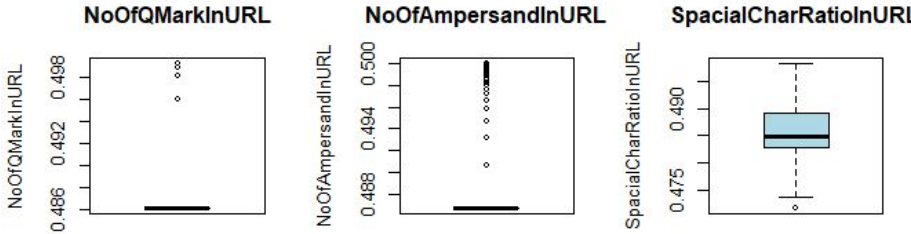
# Checking outliers







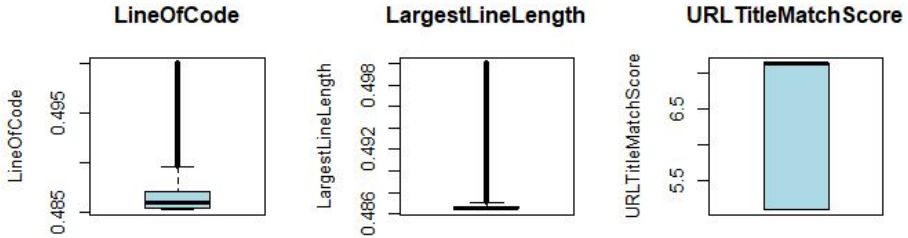






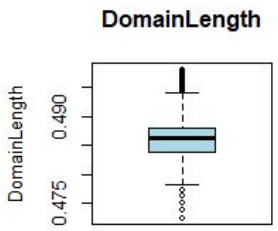


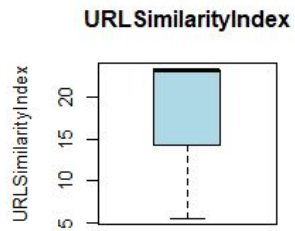



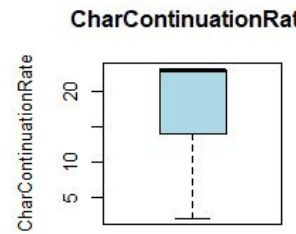


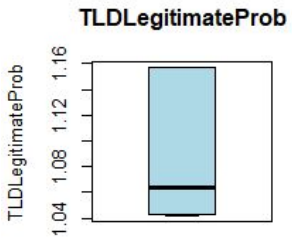


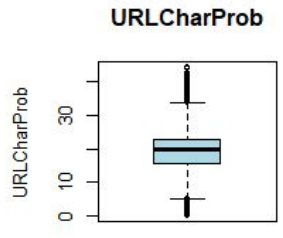


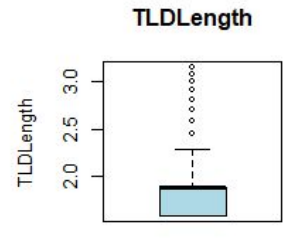



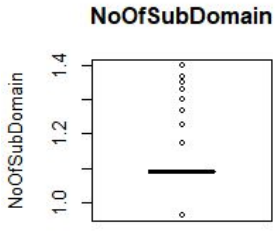



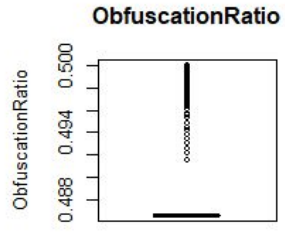



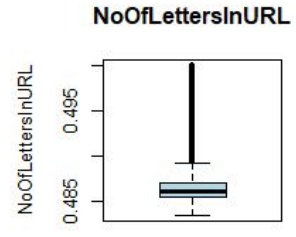



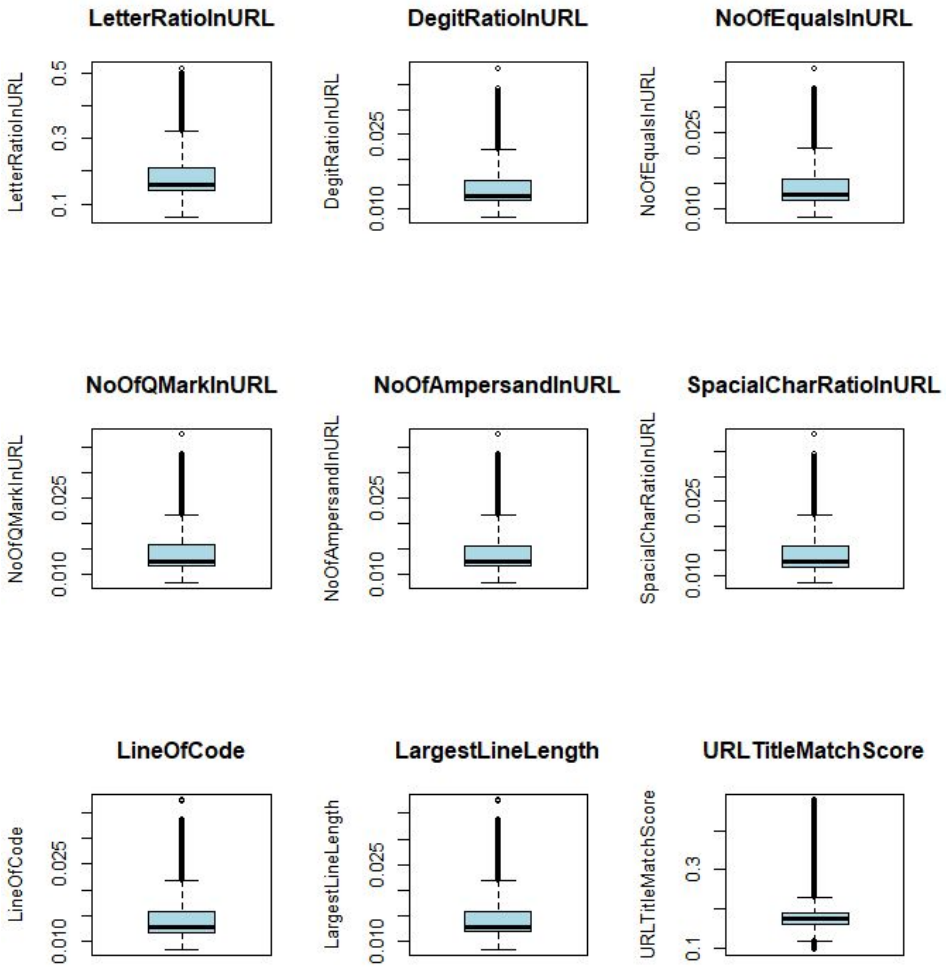



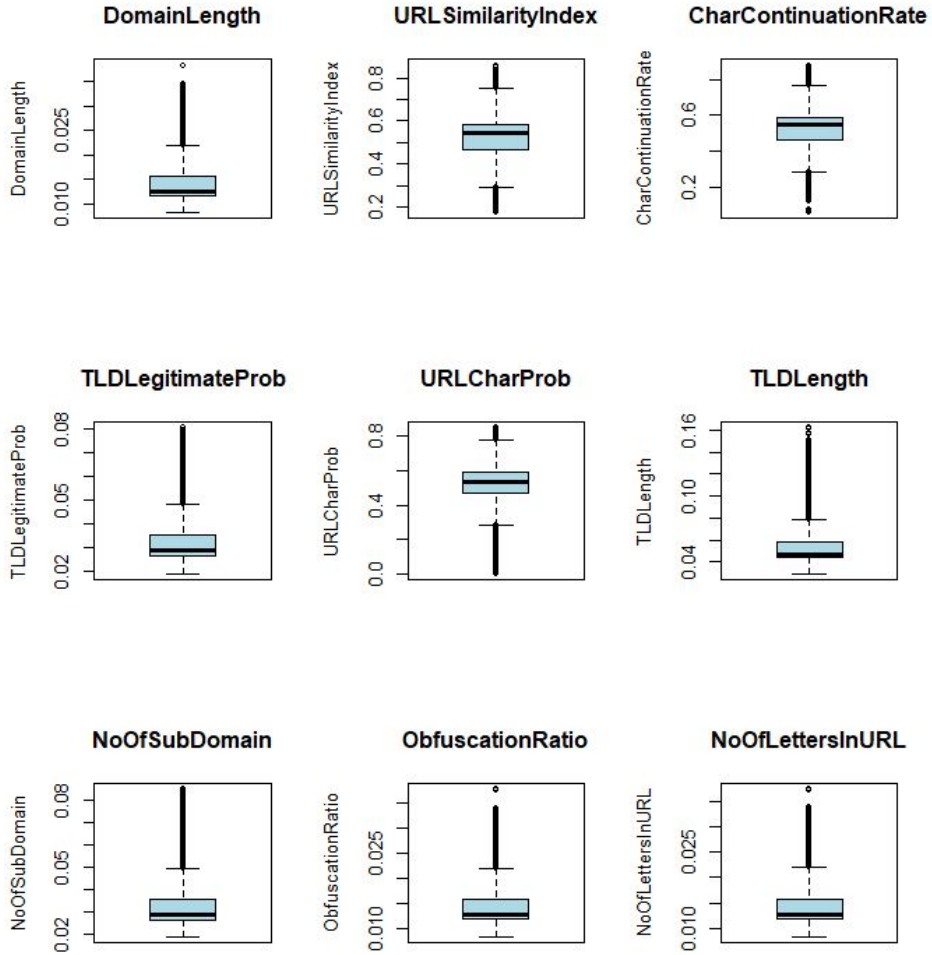



# Applied Spatial sign transformation






Michigan Tech

# Final Number of Predictors we have

- Before removing highly correlated predictors
  51(original) - 6(near zero predictors) = 49 predictors

- After removing highly correlated predictors
  51(original) -6(near zero predictors - 5(highly correlated predictors) = 44 predictors

# Classification statistic used

- Kappa is used as a classification statistics metrics, since our dataset is not perfectly balanced

# How we Spend the data?

Spend the data in to training and testing sets,

- 80/20 split using Stratified Sampling

For Resampling,

- K-fold Cross-validation(K=3)

# Model Building

## Linear Models

- Logistic Regression
- Linear Discriminant Analysis(LDA)
- Partial least squares Discriminant Analysis(PLSDA)
- Penalized model

## Non Linear Models

- Regularized Discriminant Analysis(RDA)
- Mixture Discriminant Analysis(MDA)
- Neural Networks
- Flexible Discriminant Analysis (FDA)
- Support Vector Machines (SVM)
- K- Nearest Neighbors (KNN)
- Naive Bayes

# Logistic Regression

Specific preprocessing - none

```
Generalized Linear Model

188636 samples
    49 predictor
     2 classes: 'legitimate', 'phishing'

No pre-processing
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 125757, 125758, 125757
Resampling results:

  Accuracy   Kappa
  0.8704171  0.7333645
```

```
Confusion Matrix and Statistics

            Reference
Prediction   legitimate phishing
  legitimate      26917     5475
  phishing           53    14714

               Accuracy : 0.8828
                 95% CI : (0.8798, 0.8857)
    No Information Rate : 0.5719
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7522

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9980
            Specificity : 0.7288
         Pos Pred Value : 0.8310
         Neg Pred Value : 0.9964
             Prevalence : 0.5719
         Detection Rate : 0.5708
   Detection Prevalence : 0.6869
      Balanced Accuracy : 0.8634

       'Positive' Class : legitimate
```

# LDA

Specific preprocessing - Remove highly correlated variables, Centering and scaling

```
Linear Discriminant Analysis

188636 samples
    44 predictor
     2 classes: 'legitimate', 'phishing'

Pre-processing: centered (44), scaled (44)
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 125758, 125757, 125757
Resampling results:

  Accuracy   Kappa
  0.8689752  0.7218282
```

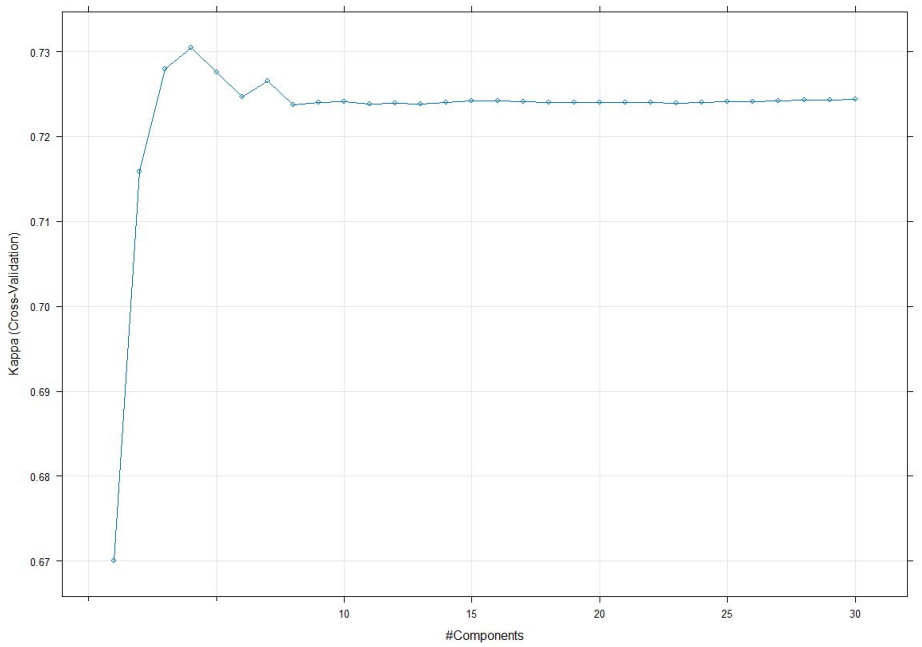
```
Confusion Matrix and Statistics

          Reference
Prediction   legitimate phishing
  legitimate      26948     6215
  phishing           22    13974

               Accuracy : 0.8677
                 95% CI : (0.8647, 0.8708)
    No Information Rate : 0.5719
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7191

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9992
            Specificity : 0.6922
         Pos Pred Value : 0.8126
         Neg Pred Value : 0.9984
             Prevalence : 0.5719
         Detection Rate : 0.5714
   Detection Prevalence : 0.7032
      Balanced Accuracy : 0.8457

       'Positive' Class : legitimate
```

# PLSDA

## Specific preprocessing - Centering and Scaling

```
Partial Least Squares

188636 samples
    49 predictor
     2 classes: 'legitimate', 'phishing'

Pre-processing: centered (49), scaled (49)
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 125757, 125758, 125757
Resampling results across tuning parameters:
```

| ncomp | Accuracy | Kappa |
|---|---|---|
| 1 | 0.8457611 | 0.6700676 |
| 2 | 0.8663087 | 0.7158344 |
| 3 | 0.8716682 | 0.7279229 |
| 4 | 0.8728186 | 0.7304395 |
| 5 | 0.8715251 | 0.7275886 |
| 6 | 0.8702051 | 0.7246423 |
| 7 | 0.8710798 | 0.7265330 |
| 8 | 0.8698287 | 0.7237495 |
| 9 | 0.8699718 | 0.7240628 |
| 10 | 0.8699877 | 0.7241075 |
| 28 | 0.8700778 | 0.7242922 |
| 29 | 0.8700619 | 0.7242562 |
| 30 | 0.8701149 | 0.7243748 |




```
Kappa was used to select the optimal model using the largest value.
The final value used for the model was ncomp = 4.
```

```
Confusion Matrix and Statistics

          Reference
Prediction   legitimate phishing
  legitimate      26927     5983
  phishing           43    14206

               Accuracy : 0.8722
                 95% CI : (0.8692, 0.8752)
    No Information Rate : 0.5719
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.729

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9984
            Specificity : 0.7037
         Pos Pred Value : 0.8182
         Neg Pred Value : 0.9970
             Prevalence : 0.5719
         Detection Rate : 0.5710
   Detection Prevalence : 0.6979
      Balanced Accuracy : 0.8510

       'Positive' Class : legitimate
```

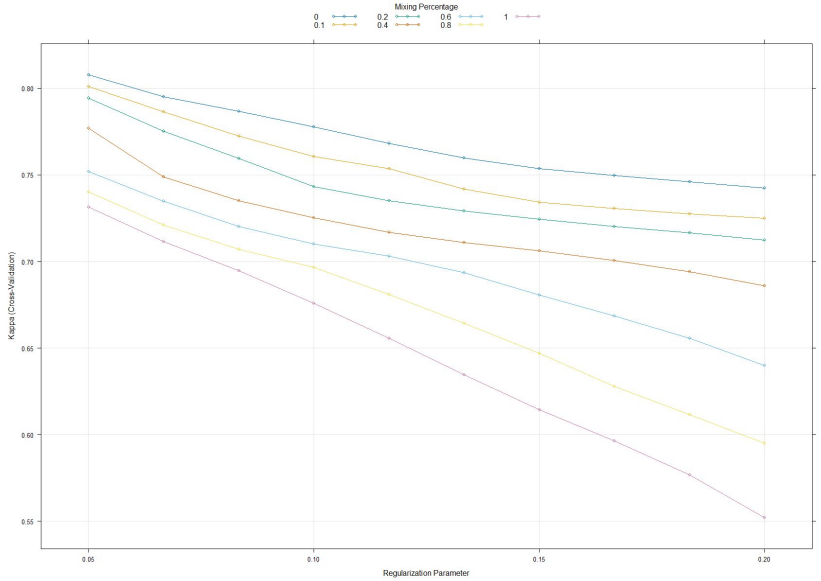
# Penalized Models

## Specific preprocessing - Centering and Scaling

```
glmnet

188636 samples
    49 predictor
     2 classes: 'legitimate', 'phishing'

Pre-processing: centered (49), scaled (49)
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 125757, 125758, 125757
Resampling results across tuning parameters:
```

| alpha | lambda | Accuracy | Kappa |
|---|---|---|---|
| 0.0 | 0.05000000 | 0.9084533 | 0.8079254 |
| 0.0 | 0.06666667 | 0.9025637 | 0.7952006 |
| 0.0 | 0.08333333 | 0.8986885 | 0.7867997 |
| 0.0 | 0.10000000 | 0.8945853 | 0.7778845 |
| 0.0 | 0.11666667 | 0.8901694 | 0.7682664 |
| 0.0 | 0.13333333 | 0.8862677 | 0.7597466 |
| 0.0 | 0.15000000 | 0.8834581 | 0.7535982 |
| 0.0 | 0.16666667 | 0.8816557 | 0.7496457 |
| 0.0 | 0.18333333 | 0.8799646 | 0.7459350 |
| 0.0 | 0.20000000 | 0.8783371 | 0.7423619 |
| 0.1 | 0.05000000 | 0.9052832 | 0.8010699 |
| 0.1 | 0.06666667 | 0.8985719 | 0.7865341 |
| 1.0 | 0.15000000 | 0.8211052 | 0.6143997 |
| 1.0 | 0.16666667 | 0.8131905 | 0.5963499 |
| 1.0 | 0.18333333 | 0.8046078 | 0.5766757 |
| 1.0 | 0.20000000 | 0.7939153 | 0.5520170 |

```
Kappa was used to select the optimal model using the largest value.
The final values used for the model were alpha = 0 and lambda = 0.05.
```




```
Confusion Matrix and Statistics

            Reference
Prediction   legitimate phishing
  legitimate      26959     4350
  phishing           11    15839

               Accuracy : 0.9075
                 95% CI : (0.9049, 0.9101)
    No Information Rate : 0.5719
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8059

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9996
            Specificity : 0.7845
         Pos Pred Value : 0.8611
         Neg Pred Value : 0.9993
             Prevalence : 0.5719
         Detection Rate : 0.5717
   Detection Prevalence : 0.6639
      Balanced Accuracy : 0.8921

       'Positive' Class : legitimate
```

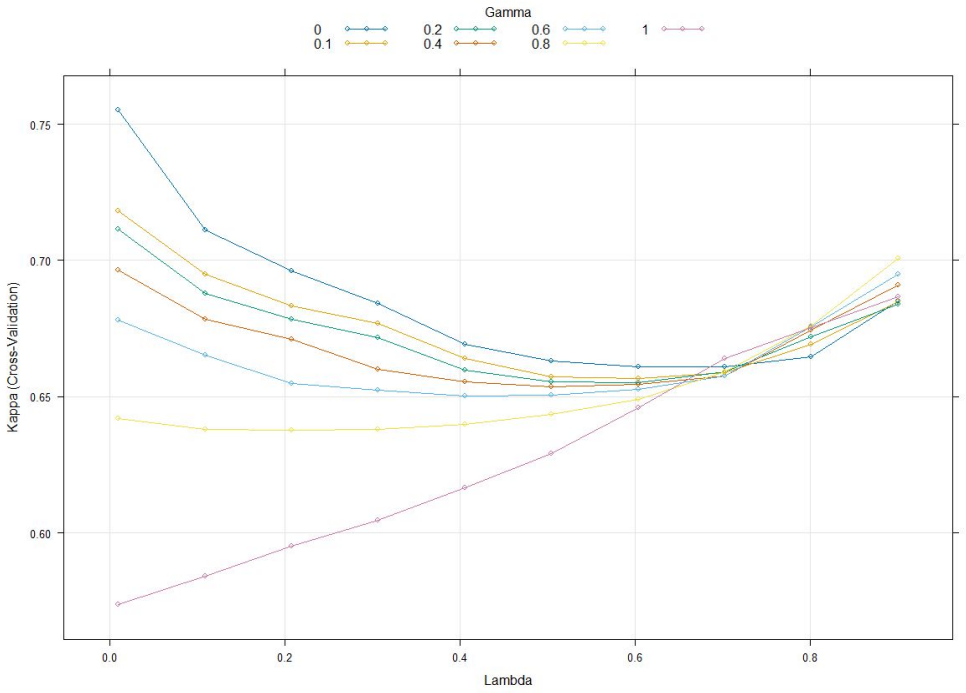
# Summary of Linear Models

| Model | Best Tuning parameter | Training Kappa | Training Accuracy | Testing Kappa | Testing Accuracy |
|---|---|---|---|---|---|
| Logistic regression | - | 0.7333645 | 0.8704171 | 0.7522 | 0.8828 |
| LDA | - | 0.7218282 | 0.8689752 | 0.7191 | 0.8677 |
| PLSDA | ncomp = 4 | 0.7304395 | 0.872818 | 0.7238 | 0.8722 |
| Penalized model | alpha = 0, lambda = 0.05 | 0.8079254 | 0.9084533 | 0.8059 | 0.9075 |

# RDA

Specific preprocessing - Remove highly correlated predictors, Center and Scale

```
Regularized Discriminant Analysis

188636 samples
    44 predictor
     2 classes: 'legitimate', 'phishing'

Pre-processing: centered (44), scaled (44)
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 125757, 125758, 125757
Resampling results across tuning parameters:

  gamma  lambda      Accuracy   Kappa
  0.0    0.0100000   0.8839246  0.7551081
  0.0    0.1088889   0.8639390  0.7110009
  0.0    0.2077778   0.8572065  0.6960433
  0.0    0.3066667   0.8518841  0.6841709
  0.0    0.4055556   0.8452257  0.6692884
  0.0    0.5044444   0.8423949  0.6629323

   1.0    0.6033333  0.8334305  0.6459694
   1.0    0.7022222  0.8420874  0.6641306
   1.0    0.8011111  0.8476378  0.6754625
   1.0    0.9000000  0.8529549  0.6865313

Kappa was used to select the optimal model using the largest value.
The final values used for the model were gamma = 0 and lambda = 0.01.
```

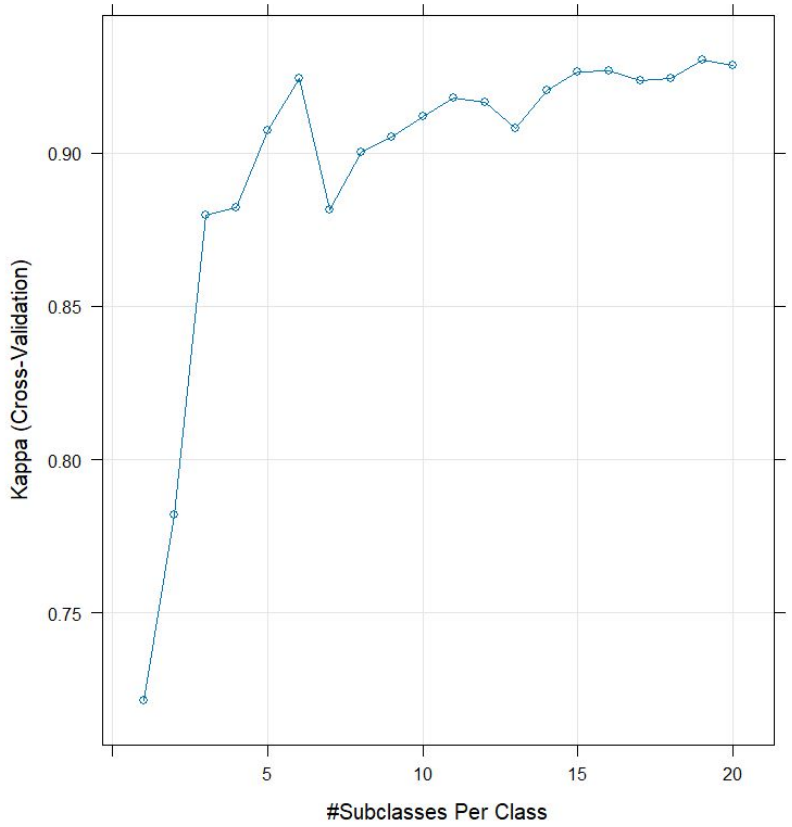



```
Confusion Matrix and Statistics

          Reference
Prediction   legitimate phishing
  legitimate      26824     5354
  phishing          146    14835

               Accuracy : 0.8834
                 95% CI : (0.8804, 0.8863)
    No Information Rate : 0.5719
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7538

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9946
            Specificity : 0.7348
         Pos Pred Value : 0.8336
         Neg Pred Value : 0.9903
             Prevalence : 0.5719
         Detection Rate : 0.5688
   Detection Prevalence : 0.6823
      Balanced Accuracy : 0.8647

       'Positive' Class : legitimate
```

# MDA

Specific preprocessing - Remove highly correlated predictors, Center and Scale

```
Mixture Discriminant Analysis

188636 samples
    44 predictor
     2 classes: 'legitimate', 'phishing'

Pre-processing: centered (44), scaled (44)
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 125757, 125758, 125757
Resampling results across tuning parameters:

  subclasses  Accuracy   Kappa
   1          0.8688426  0.7215384
   2          0.8964514  0.7820573
   3          0.9423229  0.8797886
   4          0.9434892  0.8820307
   5          0.9552525  0.9072315
   6          0.9633686  0.9244282
   7          0.9429588  0.8814820
  19          0.9661412  0.9302595
  20          0.9653884  0.9286484

Kappa was used to select the optimal model using the largest value.
The final value used for the model was subclasses = 19.
```

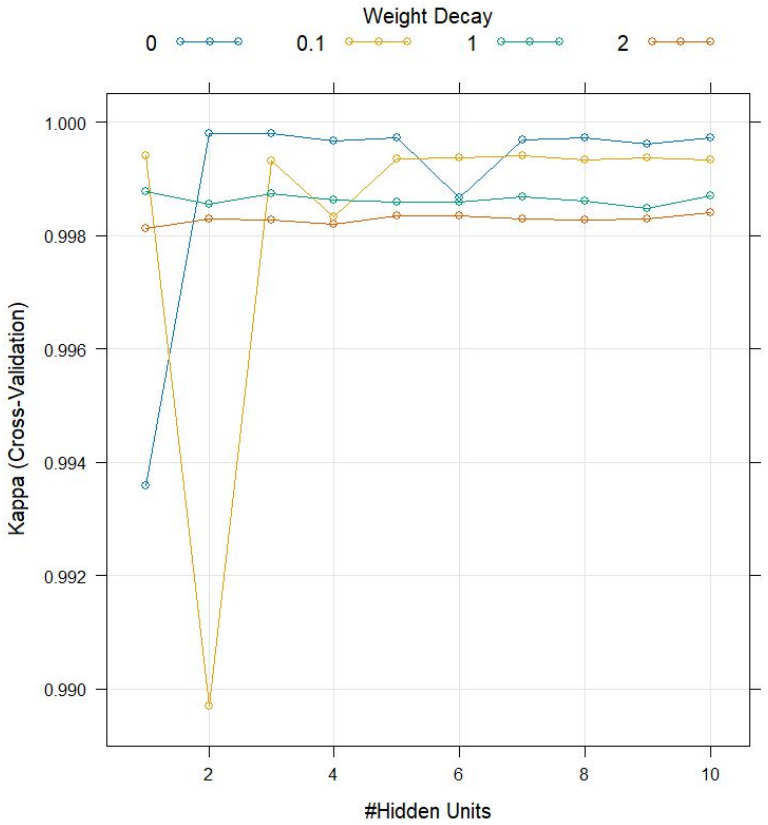



```
Confusion Matrix and Statistics

          Reference
Prediction   legitimate phishing
  legitimate      26927     1812
  phishing           43    18377

               Accuracy : 0.9607
                 95% CI : (0.9589, 0.9624)
    No Information Rate : 0.5719
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9188

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9984
            Specificity : 0.9102
         Pos Pred Value : 0.9369
         Neg Pred Value : 0.9977
             Prevalence : 0.5719
         Detection Rate : 0.5710
   Detection Prevalence : 0.6094
      Balanced Accuracy : 0.9543

       'Positive' Class : legitimate
```

# Neural Network

Specific preprocessing - Remove highly correlated predictors,Center and Scale

```
Neural Network

188636 samples
    44 predictor
     2 classes: 'legitimate', 'phishing'

Pre-processing: centered (44), scaled (44)
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 125757, 125758, 125757
Resampling results across tuning parameters:

  size  decay  Accuracy   Kappa
   1     0.0    0.9968670  0.9935898
   1     0.1    0.9997084  0.9994045
   1     1.0    0.9994010  0.9987765
   1     2.0    0.9990829  0.9981268
   2     0.0    0.9999046  0.9993051
   2     0.1    0.9949745  0.9896954
   2     1.0    0.9992896  0.9985492
   2     2.0    0.9991624  0.9982892
   3     0.0    0.9990465  0.9994051
   3     0.1    0.9996660  0.9993179
  10     1.0    0.9993639  0.9987007
  10     2.0    0.9992154  0.9983975

Kappa was used to select the optimal model using the largest value.
The final values used for the model were size = 3 and decay = 0.
```




```
Confusion Matrix and Statistics

            Reference
Prediction   legitimate phishing
  legitimate      26969        3
  phishing            1    20186

               Accuracy : 0.9996
                 95% CI : (0.9998, 1)
    No Information Rate : 0.5719
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9994

 Mcnemar's Test P-Value : 0.6171

            Sensitivity : 1.0000
            Specificity : 0.9999
         Pos Pred Value : 0.9999
         Neg Pred Value : 1.0000
             Prevalence : 0.5719
         Detection Rate : 0.5719
   Detection Prevalence : 0.5719
      Balanced Accuracy : 0.9999

       'Positive' Class : legitimate
```

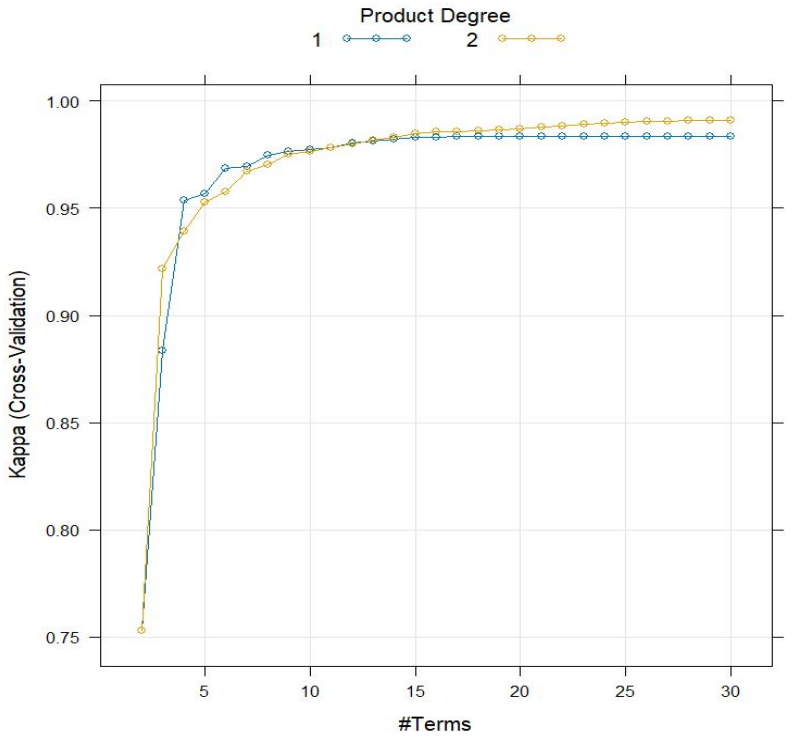
# FDA

Specific preprocessing - Remove highly correlated predictors, Spatial Sign transformation

```
Flexible Discriminant Analysis

188636 samples
    44 predictor
     2 classes: 'legitimate', 'phishing'

No pre-processing
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 125757, 125758, 125757
Resampling results across tuning parameters:
```

| degree | nprune | Accuracy | Kappa |
|---|---|---|---|
| 1 | 2 | 0.8787135 | 0.7529593 |
| 1 | 3 | 0.9431816 | 0.8834554 |
| 1 | 4 | 0.9774963 | 0.9539180 |
| 1 | 5 | 0.9789860 | 0.9569607 |
| 2 | 26 | 0.9953137 | 0.9904238 |
| 2 | 27 | 0.9953879 | 0.9905755 |
| 2 | 28 | 0.9954940 | 0.9907926 |
| 2 | 29 | 0.9955417 | 0.9908903 |
| 2 | 30 | 0.9956371 | 0.9910854 |

```
Kappa was used to select the optimal model using the largest value.
The final values used for the model were degree = 2 and nprune = 30.
```

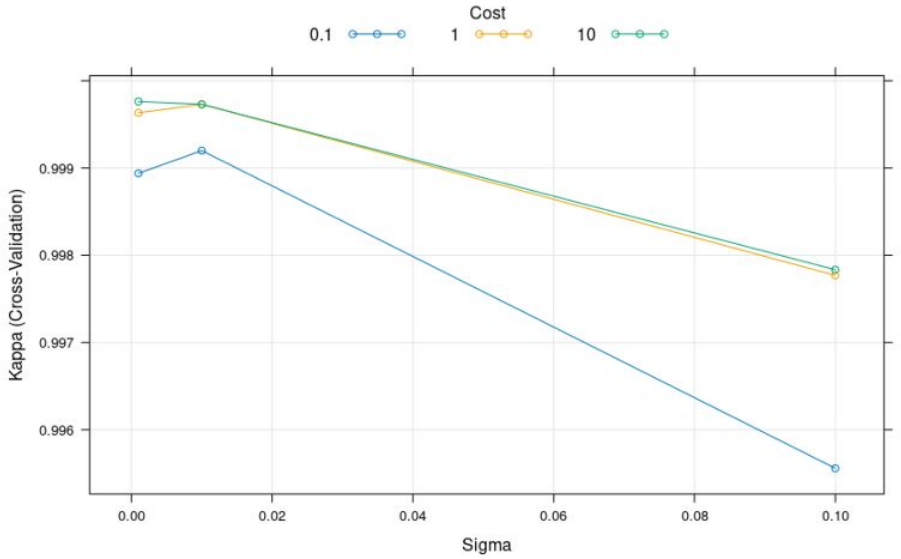



```
Confusion Matrix and Statistics

          Reference
Prediction   legitimate phishing
  legitimate      26919      143
  phishing           51    20046

               Accuracy : 0.9959
                 95% CI : (0.9953, 0.9964)
    No Information Rate : 0.5719
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9916

 Mcnemar's Test P-Value : 6.428e-11

            Sensitivity : 0.9981
            Specificity : 0.9929
         Pos Pred Value : 0.9947
         Neg Pred Value : 0.9975
             Prevalence : 0.5719
         Detection Rate : 0.5708
   Detection Prevalence : 0.5738
      Balanced Accuracy : 0.9955

       'Positive' Class : legitimate
```

# SVM

Specific preprocessing - Center and Scale

```
Support Vector Machines with Radial Basis Function Kernel

188636 samples
    49 predictor
     2 classes: 'legitimate', 'phishing'

Pre-processing: centered (49), scaled (49)
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 125757, 125758, 125757
Resampling results across tuning parameters:

  sigma  C     Accuracy   Kappa
  0.001    0.1  0.9995017  0.9989823
  0.001    1.0  0.9998516  0.9996969
  0.001   10.0  0.9998887  0.9997726
  0.100    1.0  0.9989769  0.9979099
  0.100   10.0  0.9989981  0.9979533

Kappa was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.001 and C = 10.
```

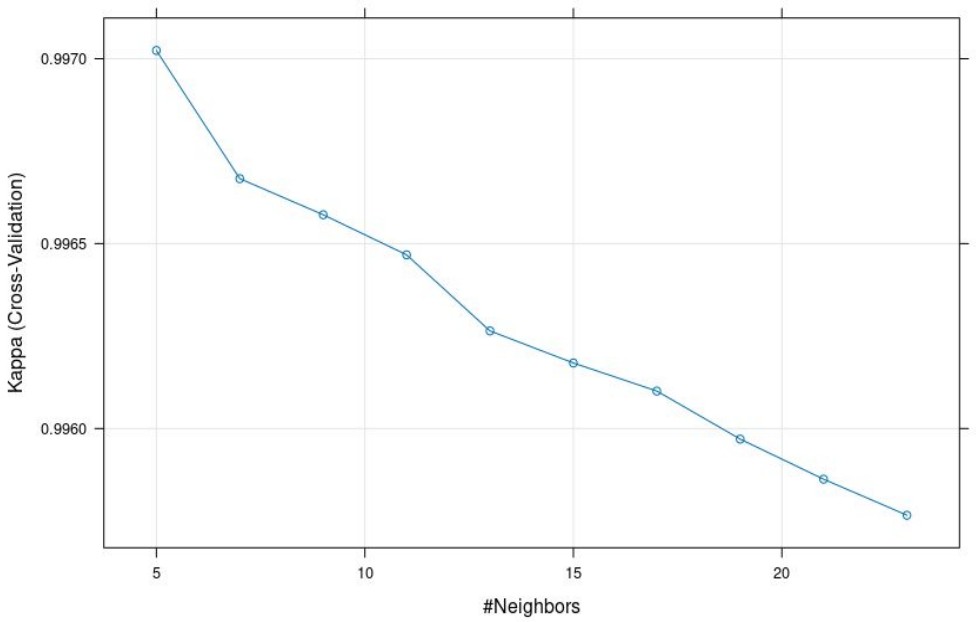



```
Confusion Matrix and Statistics

          Reference
Prediction   legitimate phishing
  legitimate      26966        5
  phishing            4    20184

               Accuracy : 0.9998
                 95% CI : (0.9996, 0.9999)
    No Information Rate : 0.5719
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9996

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9999
            Specificity : 0.9998
         Pos Pred Value : 0.9998
         Neg Pred Value : 0.9998
             Prevalence : 0.5719
         Detection Rate : 0.5718
   Detection Prevalence : 0.5719
      Balanced Accuracy : 0.9998

       'Positive' Class : legitimate
```

# KNN

Specific preprocessing - Center and Scale

```
k-Nearest Neighbors

188636 samples
    49 predictor
     2 classes: 'legitimate', 'phishing'

Pre-processing: centered (49), scaled (49)
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 125757, 125757, 125758
Resampling results across tuning parameters:

  k   Accuracy   Kappa
  5   0.9985422  0.9970222
  7   0.9983725  0.9966756
  9   0.9983248  0.9965780
 11   0.9982718  0.9964696
 13   0.9981711  0.9962639
 15   0.9981287  0.9961772
 17   0.9980916  0.9961013
 19   0.9980279  0.9959714
 21   0.9979749  0.9958631
 23   0.9979272  0.9957656

Kappa was used to select the optimal model using the largest value.
The final value used for the model was k = 5.
```




```
Confusion Matrix and Statistics

          Reference
Prediction   legitimate phishing
  legitimate      26956       44
  phishing           14    20145

               Accuracy : 0.9988
                 95% CI : (0.9984, 0.9991)
    No Information Rate : 0.5719
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9975

 Mcnemar's Test P-Value : 0.0001402

            Sensitivity : 0.9995
            Specificity : 0.9978
         Pos Pred Value : 0.9984
         Neg Pred Value : 0.9993
             Prevalence : 0.5719
         Detection Rate : 0.5716
   Detection Prevalence : 0.5725
      Balanced Accuracy : 0.9987

       'Positive' Class : legitimate
```

# Naive Bayes

Specific preprocessing - Remove highly correlated predictors, Center and Scale

```
Naive Bayes

188636 samples
    44 predictor
     2 classes: 'legitimate', 'phishing'

Pre-processing: centered (44), scaled (44)
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 125757, 125758, 125757
Resampling results across tuning parameters:

  usekernel  Accuracy   Kappa
  FALSE           NaN       NaN
   TRUE      0.9303155  0.8571713

Tuning parameter 'fL' was held constant at a value of 0
Tuning parameter 'adjust' was held constant at a value of 1
Kappa was used to select the optimal model using the largest value.
The final values used for the model were fL = 0, usekernel = TRUE and
adjust=1.
```

```
Confusion Matrix and Statistics

            Reference
Prediction   legitimate phishing
  legitimate      25688     1903
  phishing         1282    18286

               Accuracy : 0.9325
                 95% CI : (0.9302, 0.9347)
    No Information Rate : 0.5719
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8615

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9525
            Specificity : 0.9057
         Pos Pred Value : 0.9310
         Neg Pred Value : 0.9345
             Prevalence : 0.5719
         Detection Rate : 0.5447
   Detection Prevalence : 0.5851
      Balanced Accuracy : 0.9291

       'Positive' Class : legitimate
```

# Summary of Nonlinear Models

| Model | Best Tuning parameter | Training Kappa | Training Accuracy | Testing Kappa | Testing Accuracy |
|---|---|---|---|---|---|
| RDA | gamma = 0 and lambda = 0.01 | 0.7551 | 0.88392 | 0.7538 | 0.8834 |
| MDA | subclasses= 19 | 0.93025 | 0.96614 | 0.9188 | 0.9607 |
| Neural Network | size = 3 and decay = 0 | 0.999405 | 0.99904 | 0.9994 | 0.9996 |
| FDA | degree = 2 and nprune = 30 | 0.9910854 | 0.9956371 | 0.9916 | 0.9959 |
| SVM | cost=10,sigma=0.001 | 0.9997726 | 0.9998887 | 0.9996 | 0.9998 |
| KNN | k = 5 | 0.9970 | 0.99854 | 0.9975 | 0.9988 |
| Naive Bayes | - | 0.8571713 | 0.9303155 | 0.8615 | 0.9325 |

# The Best 2 Models

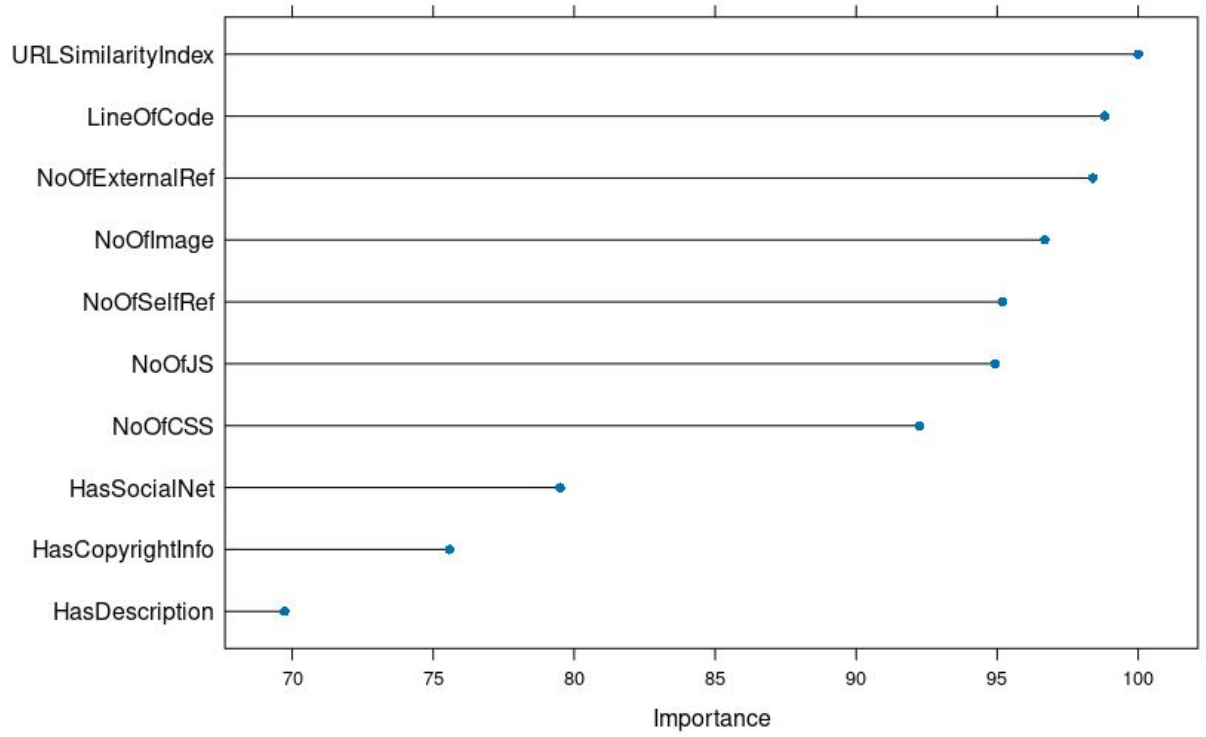
1. SVM
   Kappa - 0.9996 , Cost =10  Sigma= 0.001

2. Neural Network
   Kappa - 0.9994 ,  size=3,  decay=0

# Top Ten most Important predictors (SVM)

only 10 most important variables shown (out of 49)

| | Importance |
|---|---|
| URLSimilarityIndex | 100.00 |
| LineOfCode | 98.82 |
| NoOfExternalRef | 98.40 |
| NoOfImage | 96.69 |
| NoOfSelfRef | 95.18 |
| NoOfJS | 94.92 |
| NoOfCSS | 92.25 |
| HasSocialNet | 79.50 |
| HasCopyrightInfo | 75.59 |
| HasDescription | 69.73 |



# Lit Review & Comp: Link



**Table 4**
The comparison with previous studies.

| Publications | Task | Performance |
|---|---|---|
| [5] | phishing-website detection | Accuracy of 98.64% |
| [11] | phishing attacks detection | Accuracy of 94.612 |
| [20] | phishing URL detection on Kaggle dataset | Accuracy of 96.25% |
| [31] | phishing URL detection with PhiUSIIL dataset | Accuracy of 99.24% |
| Our study | phishing URL detection with PhiUSIIL dataset | Accuracy with 99.97% |

# Lit Review & Comp: Link



**Table 1**
Comparative summary of related work with proposed work.

| Ref. | Features | Detection model | Dataset records | Accu-racy | Limitations |
|---|---|---|---|---|---|
| Gupta et al. (2021) | URL features only | Depends on Random forest algorithm | 11964 | 99.57% | Depends solely on URL features, and on single ML algorithm, experimented on small dataset |
| Pandey and Mishra (2023) | Dominant color features and OCR | Depends on Random forest algorithm | 6200 | 99.13% | Depends solely on single ML algorithm, experimented on small dataset |
| Ahammad et al. (2022) | URL Features only | RF, DT, Light GBM, LR, and SVM | 3000 | 89.50% | Low prediction performances, experimented on small dataset |
| Jain et al. (2022) | Static and site popularity features | LR, KNN, SVM, DT, and RF | 4000 | 93.85% | Depends on third party features, low prediction performances, experimented on small dataset |
| Alani and Tawfik (2022) | URL and third-party features | RF, LR, DT, GNB, and MLP | 88646 | 97.50% | Depends on third party features, low prediction performances |
| Ding et al. (2019) | URL, HTML and third-party features | Logistic regression | 8659 | 98.90% | Depends on third party features, depends on single ML algorithm, small dataset |
| Sharma and Singh (2022) | Features from webpage HTML code | TF-IDF and AdaBoost | 50000 | 98.01% | Depends on single ML algorithm, small dataset |
| Nagunwa et al. (2022) | Features derived from DNS, host, and network | Eight ML and three DL algorithms | 11801 | 98.42% | Computationally expensive model (11 algorithms), small dataset |
| Sameen et al. (2020) | URL Features only | Boosting-based (2), Ten algorithms | 100000 | 98.00% | Computationally expensive model (10 algorithms) |
| Rao et al. (2022) | HTML code, and domain specific features | RF, SVM, LR, DT, and XGBoost | 10514 | 99.34% | Experimented on small dataset |
| Proposed | URL, HTML, and derived features | BernoulliNB, PassiveAggressive, and SGDClassifier | 235795 | 99.79% | Limitation of classifying URLs that download executable file. |

# Thank You!