Fraudulent Website Detection with Nonparametric Based Modelling

Xinyi (Vicky) Xiang   Drew Wise

Department of Statistics
University of Washington

DRP, 03/17/2022

Outline

XGBoost
- Essence of the Model
- Important Metrics Visualized

Random Forest
- CARTs based ”bootstraped” method (Hutchinson et al., 2018)

Random Forest Cont.
- Mean Decrease Gini
- Variable Importance Plot

Support Vector Machine
- In contrast with the nonparametric models above

XGBoost

Cover metric is the contribution of each feature to the number of observations summed up from each tree expressed in percentage.

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{1}$$

Gain corresponds to the importance of the node in the model.
G_L and G_R quantifies the incorrect classification at a split for the total number of classes.
As a corollary, H takes into account of the entropy from the left and right branch.

XGBoost

XGBoost advances its system with observable optimization upon the base GBM framework

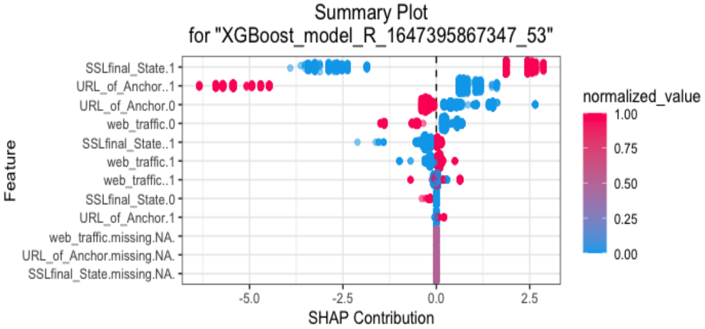


Figure: Variable Importance Heatmap from Normalized Score of Feature

Random Forest

$$\begin{aligned} \text{Gini impurity} &= 1 - \sum_{i=1}^{K} p_i^2 \\ &= 1 - \text{Gini Index} \end{aligned} \tag{2}$$

K is the number of labels, p_i is the proportion of the i^{th} label
Eval metrics in courtesy of (Subasi et al., 2017)

Random Forest Cont.

SSLfinal_state, URL_of_Anchor and web_traffic are the three most important predictors

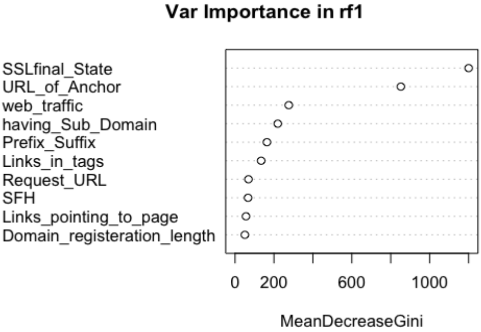
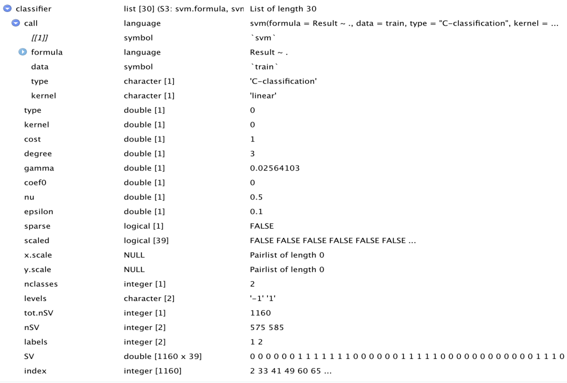


Figure: Mean Decrease Gini

SVM Classifier Parameters

Classifier parameters with 1160 support vectors



```
classifier        list [30] (S3: svm.formula, svn  List of length 30
  call            language                         svm(formula = Result ~ ., data = train, type = "C-classification", kernel = ...
    [[1]]         symbol                           `svm`
    formula       language                         Result ~ .
    data          symbol                           `train`
    type          character [1]                    'C-classification'
    kernel        character [1]                    'linear'
  type            double [1]                       0
  kernel          double [1]                       0
  cost            double [1]                       1
  degree          double [1]                       3
  gamma           double [1]                       0.02564103
  coef0           double [1]                       0
  nu              double [1]                       0.5
  epsilon         double [1]                       0.1
  sparse          logical [1]                      FALSE
  scaled          logical [39]                     FALSE FALSE FALSE FALSE FALSE FALSE ...
  x.scale         NULL                             Pairlist of length 0
  y.scale         NULL                             Pairlist of length 0
  nclasses        integer [1]                      2
  levels          character [2]                    '-1' '1'
  tot.nSV         integer [1]                      1160
  nSV             integer [2]                      575 585
  labels          integer [2]                      1 2
  SV              double [1160 x 39]               0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0
  index           integer [1160]                   2 33 41 49 60 65 ...
```

Figure: SVM Classifier Parameters

Thank you!

xinyix7@uw.edu

References

Hutchinson, S., Zhang, Z., and Liu, Q. (2018). *Detecting Phishing Websites with Random Forest: Third International Conference, MLICOM 2018, Hangzhou, China, July 6-8, 2018, Proceedings*, pages 470–479.

Subasi, A., Molah, E., Almkallawi, F., and Chaudhery, T. J. (2017). Intelligent phishing website detection using random forest classifier. pages 1–5.