Random Forest Variable Importance and Variable Selection

Abstract

Introduction

Variable Importance

Variable Selection

Testing Methodology

To test out all different R packages that implement variable importance and variable selection methods, credit default dataset composed by Professor J.F.Plante for Statistical Learning class will be used as a sample dataset. For testing purposes, the original dataset was undersampled from 1 million rows to 10000 rows while maintaining only complete rows with preserving the original ratio of target binary variable that defines whether the client is going to default or not. This dataset will help to benchmark various R packages and find similarities and difference between their respective methods on random forest's variable importance and variable selection. The structure of the dataset is as following:

R packages tutorials

randomForest R package

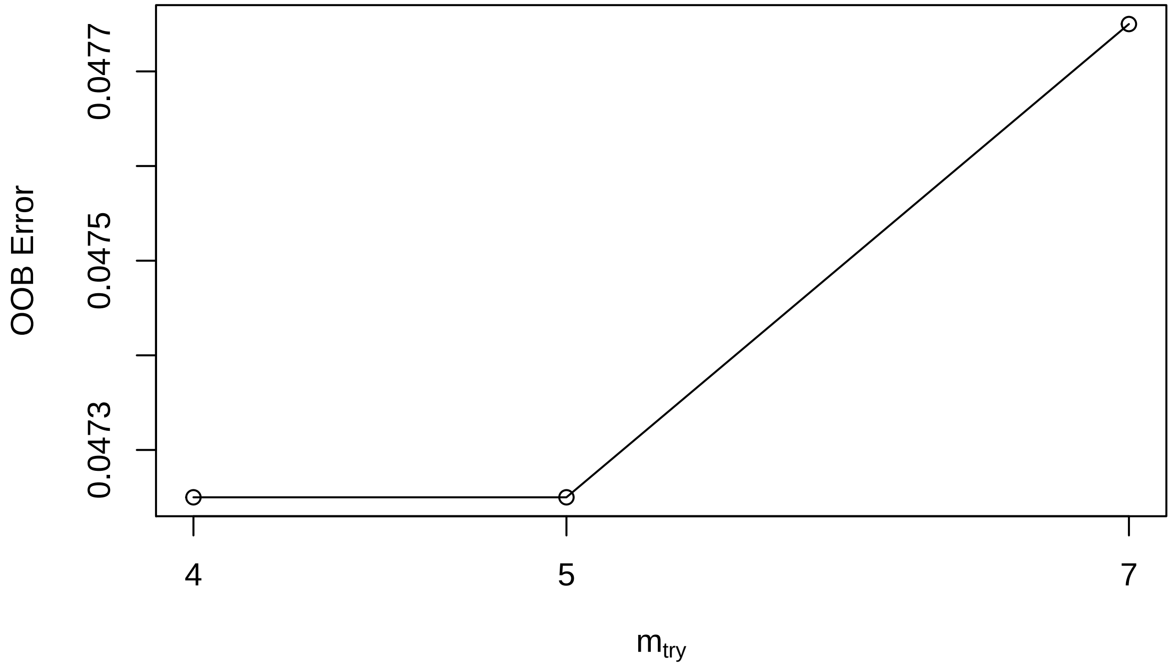
randomForest R package is the main package that implements random forest decision tree model in R. This R package is not only able to fit a random forest model, but also has built-in functions to derive variable importance for the fitted model such as *importance* and *varImpPlot*. While these functions will show variable importance, it is also vital to tune random forest hyper-parameters such as *mtry* and *ntrees* to get relevant and optimal variable importance for the optimal model. For this purpose, *tuneRF* function performs hyper-parameter search. The parameters of *tuneRF* are as following.

Parameter	Description
x	Data frame of predictors
y	Response label vector
mtryStart	Starting number for mtry
ntreeTry	Number of trees per tuning step
stepFactor	multiplier for mtry parameter for next iteration
improve	minimum OOB error improvement to continue search
plot	plot the mtry to OOB error graph
trace	output error and mtry per each iteration
doBest	run Random Forest fit based on best mtry

For the test dataset, the following hyper-parameters were used. As a result, the optimal value for mtry is 4 while number of trees were set to 500.

```r
tuneRF(train_set[,2:29],train_set[,30], stepFactor = 1.5,
       plot=TRUE, trace=TRUE, doBest=FALSE)
```

```
## mtry = 5  OOB error = 4.72%
## Searching left ...
## mtry = 4     OOB error = 4.72%
## 0 0.05
## Searching right ...
## mtry = 7     OOB error = 4.78%
## -0.01058201 0.05
```



```
##       mtry OOBError
## 4.OOB    4  0.04725
## 5.OOB    5  0.04725
## 7.OOB    7  0.04775
```

importance function

varImpPlot function shows the plot of all variables used in RandomForest model fit and their respective importance on MSE % increase, from which the statistician can derive each variable's importance on random forest model fit.

```
##             MeanDecreaseAccuracy
## NB_EMPT                2.8699425
## R_ATD                 18.5691354
## DUREE                  4.6997891
## PRT_VAL                6.7678740
## AGE_D                  8.4082814
## REV_BT                19.3692975
## REV_NET               20.1139236
## TYP_RES                5.0777388
## ST_EMPL               -1.0306430
## MNT_EPAR              23.1489588
## NB_ER_6MS                13.4172564
## NB_ER_12MS               15.4306215
## NB_DEC_12MS               9.9702681
## NB_OPER                  15.7417438
## NB_COUR                  10.8061117
## NB_INTR_1M                0.1128450
## NB_INTR_12M              -0.7673524
## PIR_DEL                   9.4078365
## NB_DEL_30                 4.6811041
## NB_DEL_60                 6.3915120
## NB_DEL_90                13.2756341
## MNT_PASS                 14.9102728
## MNT_ACT                  23.3376112
## MNT_AUT_REN              20.0796449
## MNT_UTIL_REN             19.8234313
## NB_SATI                  11.5441295
## TYP_FIN                   0.0000000
## MNT_DEMANDE               0.3842530
```

Conclusion

References