Workshop 4.9 Solution

Veerasak Kritsanapraphan

2/1/2020

library(class)  
library(caret)  
library(e1071)  
library(adabag)  
library(gbm)  
library(fastDummies)  
library(randomForest)  
library(dplyr)  
telco <- read.csv("WA\_Fn-UseC\_-Telco-Customer-Churn.csv", header=TRUE)

### Remove NA

str(telco)

## 'data.frame': 7043 obs. of 21 variables:  
## $ customerID : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 2565 5536 6512 6552 1003 4771 5605 4535 ...  
## $ gender : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...  
## $ SeniorCitizen : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Partner : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...  
## $ Dependents : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...  
## $ tenure : int 1 34 2 45 2 8 22 10 28 62 ...  
## $ PhoneService : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...  
## $ MultipleLines : Factor w/ 3 levels "No","No phone service",..: 2 1 1 2 1 3 3 2 3 1 ...  
## $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...  
## $ OnlineSecurity : Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 1 3 1 3 ...  
## $ OnlineBackup : Factor w/ 3 levels "No","No internet service",..: 3 1 3 1 1 1 3 1 1 3 ...  
## $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 1 1 3 1 ...  
## $ TechSupport : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 1 1 3 1 ...  
## $ StreamingTV : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 3 1 3 1 ...  
## $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 1 1 3 1 ...  
## $ Contract : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...  
## $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...  
## $ PaymentMethod : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...  
## $ MonthlyCharges : num 29.9 57 53.9 42.3 70.7 ...  
## $ TotalCharges : num 29.9 1889.5 108.2 1840.8 151.7 ...  
## $ Churn : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...

telco <- na.omit(telco)

### Data Sampling

set.seed(123)  
index <- sample(2, nrow(telco), replace=TRUE, prob=c(0.7,0.3) )  
traindata <- telco[index==1,]  
testdata <- telco[index==2,]  
sprintf("Number of Record in Training Dataset is %d" , nrow(traindata))

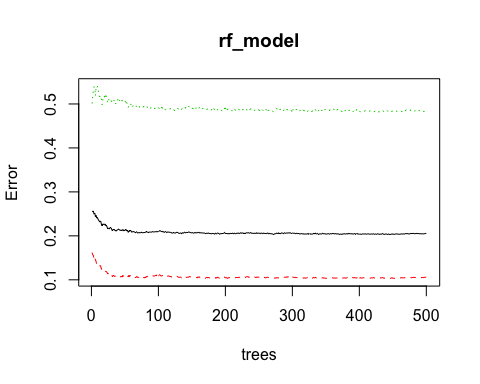
## [1] "Number of Record in Training Dataset is 4943"

sprintf("Number of Record in Testing Dataset is %d" , nrow(testdata))

## [1] "Number of Record in Testing Dataset is 2089"

### Train Random Forest Model

Formula <- Churn ~ gender + SeniorCitizen + Partner + tenure + PhoneService + MultipleLines + OnlineBackup + Contract + PaperlessBilling + PaymentMethod + MonthlyCharges + TotalCharges  
  
rf\_model <- randomForest(Formula, data=traindata, nTree=500)   
plot(rf\_model)



### Evaluate Testing Model

prediction <- predict(rf\_model, testdata)  
confusionMatrix(prediction, testdata$Churn)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 1372 286  
## Yes 148 283  
##   
## Accuracy : 0.7922   
## 95% CI : (0.7742, 0.8095)  
## No Information Rate : 0.7276   
## P-Value [Acc > NIR] : 5.467e-12   
##   
## Kappa : 0.4328   
##   
## Mcnemar's Test P-Value : 4.826e-11   
##   
## Sensitivity : 0.9026   
## Specificity : 0.4974   
## Pos Pred Value : 0.8275   
## Neg Pred Value : 0.6566   
## Prevalence : 0.7276   
## Detection Rate : 0.6568   
## Detection Prevalence : 0.7937   
## Balanced Accuracy : 0.7000   
##   
## 'Positive' Class : No   
##

### Importance Valiables

#getTree(rf\_model, 1, labelVar=TRUE)  
importance(rf\_model)

## MeanDecreaseGini  
## gender 40.971503  
## SeniorCitizen 36.777908  
## Partner 37.226718  
## tenure 301.092849  
## PhoneService 9.134068  
## MultipleLines 44.591606  
## OnlineBackup 77.218328  
## Contract 173.373595  
## PaperlessBilling 45.381730  
## PaymentMethod 132.205735  
## MonthlyCharges 351.219441  
## TotalCharges 326.452551

## Train Boosting Model

churn.adaboost <- boosting(Formula, data=traindata, boost=TRUE, mfinal=10)

### Predict

table(churn.adaboost$class, traindata$Churn)

##   
## No Yes  
## No 3274 575  
## Yes 369 725

prediction <- predict(churn.adaboost, newdata=testdata)  
table(prediction$class, testdata$Churn)

##   
## No Yes  
## No 1369 261  
## Yes 151 308

confusionMatrix(as.factor(prediction$class), testdata$Churn)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 1369 261  
## Yes 151 308  
##   
## Accuracy : 0.8028   
## 95% CI : (0.7851, 0.8196)  
## No Information Rate : 0.7276   
## P-Value [Acc > NIR] : 9.509e-16   
##   
## Kappa : 0.4704   
##   
## Mcnemar's Test P-Value : 7.872e-08   
##   
## Sensitivity : 0.9007   
## Specificity : 0.5413   
## Pos Pred Value : 0.8399   
## Neg Pred Value : 0.6710   
## Prevalence : 0.7276   
## Detection Rate : 0.6553   
## Detection Prevalence : 0.7803   
## Balanced Accuracy : 0.7210   
##   
## 'Positive' Class : No   
##

### Variable Importance

importanceplot(churn.adaboost, horiz=TRUE, cex.names=.6)

