workshop 4.2 solution

Veerasak Kritsanapraphan

1/22/2020

library(party)  
telco <- read.csv("WA\_Fn-UseC\_-Telco-Customer-Churn.csv", header=TRUE)

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 ...

## Convert Data

### Data Sampling

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 4914"

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

## [1] "Number of Record in Testing Dataset is 2129"

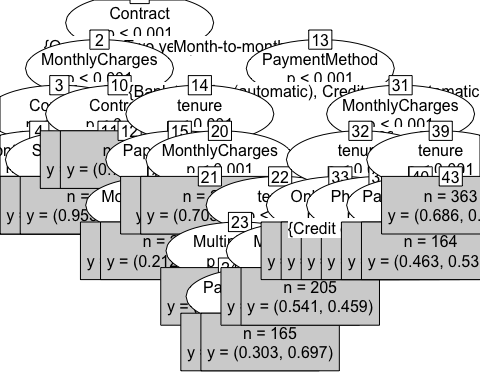
### Train Model

formular <- Churn ~ gender + SeniorCitizen + Partner + tenure + PhoneService + MultipleLines + OnlineBackup + Contract + PaperlessBilling + PaymentMethod + MonthlyCharges + TotalCharges  
ctree\_model <- ctree(formular, data = traindata)  
print(ctree\_model)

##   
## Conditional inference tree with 22 terminal nodes  
##   
## Response: Churn   
## Inputs: gender, SeniorCitizen, Partner, tenure, PhoneService, MultipleLines, OnlineBackup, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges   
## Number of observations: 4914   
##   
## 1) Contract == {One year, Two year}; criterion = 1, statistic = 803.163  
## 2) MonthlyCharges <= 93.65; criterion = 1, statistic = 77.959  
## 3) Contract == {One year}; criterion = 1, statistic = 47.056  
## 4) MonthlyCharges <= 24.85; criterion = 0.995, statistic = 12.378  
## 5)\* weights = 243   
## 4) MonthlyCharges > 24.85  
## 6)\* weights = 522   
## 3) Contract == {Two year}  
## 7) SeniorCitizen <= 0; criterion = 0.955, statistic = 8.367  
## 8)\* weights = 862   
## 7) SeniorCitizen > 0  
## 9)\* weights = 64   
## 2) MonthlyCharges > 93.65  
## 10) Contract == {Two year}; criterion = 1, statistic = 18.524  
## 11)\* weights = 246   
## 10) Contract == {One year}  
## 12)\* weights = 283   
## 1) Contract == {Month-to-month}  
## 13) PaymentMethod == {Electronic check}; criterion = 1, statistic = 120.626  
## 14) tenure <= 5; criterion = 1, statistic = 67.696  
## 15) PaperlessBilling == {Yes}; criterion = 1, statistic = 20.779  
## 16) MonthlyCharges <= 50.15; criterion = 0.966, statistic = 8.868  
## 17)\* weights = 63   
## 16) MonthlyCharges > 50.15  
## 18)\* weights = 255   
## 15) PaperlessBilling == {No}  
## 19)\* weights = 134   
## 14) tenure > 5  
## 20) MonthlyCharges <= 74.25; criterion = 1, statistic = 28.452  
## 21)\* weights = 241   
## 20) MonthlyCharges > 74.25  
## 22) tenure <= 28; criterion = 1, statistic = 17.561  
## 23) MultipleLines == {No}; criterion = 0.959, statistic = 8.523  
## 24)\* weights = 128   
## 23) MultipleLines == {Yes}  
## 25) PaperlessBilling == {No}; criterion = 0.96, statistic = 8.579  
## 26)\* weights = 27   
## 25) PaperlessBilling == {Yes}  
## 27)\* weights = 165   
## 22) tenure > 28  
## 28) MultipleLines == {No}; criterion = 0.982, statistic = 10.018  
## 29)\* weights = 48   
## 28) MultipleLines == {Yes}  
## 30)\* weights = 205   
## 13) PaymentMethod == {Bank transfer (automatic), Credit card (automatic), Mailed check}  
## 31) MonthlyCharges <= 69.7; criterion = 1, statistic = 43.617  
## 32) tenure <= 5; criterion = 1, statistic = 37.792  
## 33) OnlineBackup == {No internet service}; criterion = 0.998, statistic = 17.001  
## 34)\* weights = 179   
## 33) OnlineBackup == {No, Yes}  
## 35)\* weights = 195   
## 32) tenure > 5  
## 36) PhoneService == {Yes}; criterion = 0.976, statistic = 9.62  
## 37)\* weights = 377   
## 36) PhoneService == {No}  
## 38)\* weights = 89   
## 31) MonthlyCharges > 69.7  
## 39) tenure <= 15; criterion = 1, statistic = 53.571  
## 40) PaymentMethod == {Bank transfer (automatic)}; criterion = 0.984, statistic = 13.245  
## 41)\* weights = 61   
## 40) PaymentMethod == {Credit card (automatic), Mailed check}  
## 42)\* weights = 164   
## 39) tenure > 15  
## 43)\* weights = 363

### Plot Model

plot(ctree\_model, type="simple")



### Evaluate Model

confusionmatrix <- table(predict(ctree\_model), traindata$Churn)  
confusionmatrix

##   
## No Yes  
## No 3346 726  
## Yes 281 561

(confusionmatrix[1,1]+confusionmatrix[2,2])/  
 (confusionmatrix[1,1]+confusionmatrix[2,2]+confusionmatrix[1,2]+confusionmatrix[2,1])

## [1] 0.7950753

### Stratify Sample

library(fifer)

## Loading required package: MASS

stratefytelco <- stratified(telco, "Churn", size = c(1500, 1500))

## 'size' vector entered as:  
##   
## size = structure(c(1500, 1500),  
## .Names = c('No', 'Yes'))

table(stratefytelco$Churn)

##   
## No Yes   
## 1500 1500

### Data Sampling

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

## [1] "Number of Record in Training Dataset is 2103"

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

## [1] "Number of Record in Testing Dataset is 897"

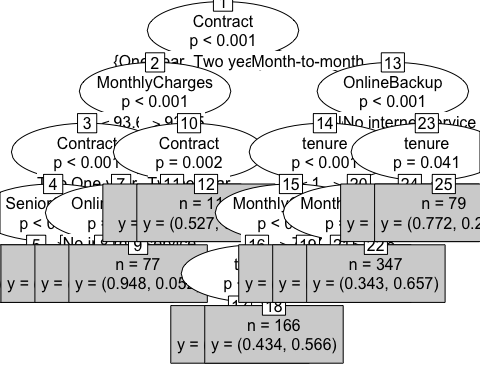
### Train Model

formular <- Churn ~ gender + SeniorCitizen + Partner + tenure + PhoneService + MultipleLines + OnlineBackup + Contract + PaperlessBilling + PaymentMethod + MonthlyCharges + TotalCharges  
ctree\_model <- ctree(formular, data = traindata)  
print(ctree\_model)

##   
## Conditional inference tree with 13 terminal nodes  
##   
## Response: Churn   
## Inputs: gender, SeniorCitizen, Partner, tenure, PhoneService, MultipleLines, OnlineBackup, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges   
## Number of observations: 2103   
##   
## 1) Contract == {One year, Two year}; criterion = 1, statistic = 530.123  
## 2) MonthlyCharges <= 93.65; criterion = 1, statistic = 61.711  
## 3) Contract == {Two year}; criterion = 1, statistic = 28.719  
## 4) SeniorCitizen <= 0; criterion = 1, statistic = 18.167  
## 5)\* weights = 265   
## 4) SeniorCitizen > 0  
## 6)\* weights = 14   
## 3) Contract == {One year}  
## 7) OnlineBackup == {No, Yes}; criterion = 0.956, statistic = 11.179  
## 8)\* weights = 152   
## 7) OnlineBackup == {No internet service}  
## 9)\* weights = 77   
## 2) MonthlyCharges > 93.65  
## 10) Contract == {Two year}; criterion = 0.998, statistic = 13.894  
## 11)\* weights = 81   
## 10) Contract == {One year}  
## 12)\* weights = 110   
## 1) Contract == {Month-to-month}  
## 13) OnlineBackup == {No, Yes}; criterion = 1, statistic = 87.398  
## 14) tenure <= 16; criterion = 1, statistic = 64.772  
## 15) MonthlyCharges <= 78.75; criterion = 1, statistic = 30.116  
## 16) tenure <= 4; criterion = 1, statistic = 23.565  
## 17)\* weights = 292   
## 16) tenure > 4  
## 18)\* weights = 166   
## 15) MonthlyCharges > 78.75  
## 19)\* weights = 307   
## 14) tenure > 16  
## 20) MonthlyCharges <= 74.8; criterion = 0.999, statistic = 15.477  
## 21)\* weights = 136   
## 20) MonthlyCharges > 74.8  
## 22)\* weights = 347   
## 13) OnlineBackup == {No internet service}  
## 23) tenure <= 3; criterion = 0.959, statistic = 8.534  
## 24)\* weights = 77   
## 23) tenure > 3  
## 25)\* weights = 79

### Plot Model

plot(ctree\_model, type="simple")



### Evaluate Model

confusionmatrix <- table(predict(ctree\_model), traindata$Churn)  
confusionmatrix

##   
## No Yes  
## No 720 194  
## Yes 308 881

(confusionmatrix[1,1]+confusionmatrix[2,2])/  
 (confusionmatrix[1,1]+confusionmatrix[2,2]+confusionmatrix[1,2]+confusionmatrix[2,1])

## [1] 0.7612934