workshop 4.6 solution

Veerasak Kritsanapraphan

1/22/2020

library(class)  
library(caret)  
library(e1071)  
library(fastDummies)  
library(dplyr)  
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

### Dummy Column

telcodum <- select(telco, Churn, gender,SeniorCitizen,Partner,tenure,PhoneService, MultipleLines,OnlineBackup,Contract,PaperlessBilling,PaymentMethod, MonthlyCharges ,TotalCharges)  
telcodum <- dummy\_cols(telcodum)

### Data Sampling

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

## [1] "Number of Record in Training Dataset is 4918"

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

## [1] "Number of Record in Testing Dataset is 2125"

### Train Model

output\_telco\_train <- traindata[,'Churn']  
  
input\_telco\_train <- traindata[, c(12,36)]  
  
input\_telco\_test <- testdata[, c(12,36)]   
  
output\_telco\_test <- testdata[,'Churn']  
  
prediction <- knn(input\_telco\_train, input\_telco\_test, output\_telco\_train, k=5)

### Evaluate Testing Model

is.factor(prediction)

## [1] TRUE

output\_telco\_test <- as.factor(output\_telco\_test)  
is.factor(output\_telco\_test)

## [1] TRUE

confusionMatrix(prediction, output\_telco\_test)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 1386 450  
## Yes 163 126  
##   
## Accuracy : 0.7115   
## 95% CI : (0.6917, 0.7307)  
## No Information Rate : 0.7289   
## P-Value [Acc > NIR] : 0.9657   
##   
## Kappa : 0.1346   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.8948   
## Specificity : 0.2188   
## Pos Pred Value : 0.7549   
## Neg Pred Value : 0.4360   
## Prevalence : 0.7289   
## Detection Rate : 0.6522   
## Detection Prevalence : 0.8640   
## Balanced Accuracy : 0.5568   
##   
## 'Positive' Class : No   
##