workshop 4.5 solution

Veerasak Kritsanapraphan

1/22/2020

library(e1071)  
library(caret)  
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 4937"

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

## [1] "Number of Record in Testing Dataset is 2106"

### Train Model

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

##   
## Naive Bayes Classifier for Discrete Predictors  
##   
## Call:  
## naiveBayes.default(x = X, y = Y, laplace = laplace)  
##   
## A-priori probabilities:  
## Y  
## No Yes   
## 0.7336439 0.2663561   
##   
## Conditional probabilities:  
## gender  
## Y Female Male  
## No 0.4889564 0.5110436  
## Yes 0.5148289 0.4851711  
##   
## SeniorCitizen  
## Y [,1] [,2]  
## No 0.1336278 0.3402990  
## Yes 0.2555133 0.4363153  
##   
## Partner  
## Y No Yes  
## No 0.4679735 0.5320265  
## Yes 0.6463878 0.3536122  
##   
## tenure  
## Y [,1] [,2]  
## No 37.58559 24.08215  
## Yes 18.14981 19.70972  
##   
## PhoneService  
## Y No Yes  
## No 0.10215351 0.89784649  
## Yes 0.09429658 0.90570342  
##   
## MultipleLines  
## Y No No phone service Yes  
## No 0.49282165 0.10215351 0.40502485  
## Yes 0.44942966 0.09429658 0.45627376  
##   
## OnlineBackup  
## Y No No internet service Yes  
## No 0.3660961 0.2747101 0.3591938  
## Yes 0.6555133 0.0608365 0.2836502  
##   
## Contract  
## Y Month-to-month One year Two year  
## No 0.43014909 0.25648813 0.31336278  
## Yes 0.87984791 0.09809886 0.02205323  
##   
## PaperlessBilling  
## Y No Yes  
## No 0.4632800 0.5367200  
## Yes 0.2319392 0.7680608  
##   
## PaymentMethod  
## Y Bank transfer (automatic) Credit card (automatic) Electronic check  
## No 0.2468250 0.2523468 0.2515185  
## Yes 0.1307985 0.1178707 0.5931559  
## PaymentMethod  
## Y Mailed check  
## No 0.2493098  
## Yes 0.1581749  
##   
## MonthlyCharges  
## Y [,1] [,2]  
## No 61.02296 31.21046  
## Yes 74.64118 24.86590  
##   
## TotalCharges  
## Y [,1] [,2]  
## No 2547.220 2336.167  
## Yes 1561.543 1921.926

### Evaluate Model

prediction <- predict(bayes\_model, newdata=traindata)  
confusionMatrix(prediction, traindata$Churn)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 2826 373  
## Yes 796 942  
##   
## Accuracy : 0.7632   
## 95% CI : (0.7511, 0.775)  
## No Information Rate : 0.7336   
## P-Value [Acc > NIR] : 1.067e-06   
##   
## Kappa : 0.4504   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.7802   
## Specificity : 0.7163   
## Pos Pred Value : 0.8834   
## Neg Pred Value : 0.5420   
## Prevalence : 0.7336   
## Detection Rate : 0.5724   
## Detection Prevalence : 0.6480   
## Balanced Accuracy : 0.7483   
##   
## 'Positive' Class : No   
##

### Evaluate Testing Model

prediction <- predict(bayes\_model, newdata=testdata)  
confusionMatrix(prediction, testdata$Churn)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 1205 152  
## Yes 347 402  
##   
## Accuracy : 0.7631   
## 95% CI : (0.7443, 0.7811)  
## No Information Rate : 0.7369   
## P-Value [Acc > NIR] : 0.003232   
##   
## Kappa : 0.451   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.7764   
## Specificity : 0.7256   
## Pos Pred Value : 0.8880   
## Neg Pred Value : 0.5367   
## Prevalence : 0.7369   
## Detection Rate : 0.5722   
## Detection Prevalence : 0.6443   
## Balanced Accuracy : 0.7510   
##   
## 'Positive' Class : No   
##