exam.R

B0113

2020-01-07

wt <- read.csv("weather\_data.csv")  
str(wt)

## 'data.frame': 366 obs. of 11 variables:  
## $ tempmin : num 8 14 13.7 13.3 7.6 6.2 6.1 8.3 8.8 8.4 ...  
## $ tempmax : num 24.3 26.9 23.4 15.5 16.1 16.9 18.2 17 19.5 22.8 ...  
## $ rainfall : num 0 3.6 3.6 39.8 2.8 0 0.2 0 0 16.2 ...  
## $ sunshine : num 6.3 9.7 3.3 9.1 10.6 8.2 8.4 4.6 4.1 7.7 ...  
## $ windgustspeed: int 30 39 85 54 50 44 43 41 48 31 ...  
## $ windspeed : int 20 17 6 24 28 24 26 24 17 6 ...  
## $ humidity : int 29 36 69 56 49 57 47 57 48 32 ...  
## $ pressure : num 1015 1008 1007 1007 1018 ...  
## $ cloud : int 7 3 7 7 7 5 6 7 7 1 ...  
## $ raintomorrow : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 1 1 1 2 1 ...  
## $ xgrp : int 1 1 1 1 1 1 1 1 1 1 ...

### 1  
# raintomorrow 컬럼은 "Yes"를 1로, "No"를 0으로 변경  
wt$raintomorrow <- ifelse(wt$raintomorrow == "Yes", 1, 0)  
  
# 누락값 인덱스 함수 만들기  
na\_idx <- function(x) {  
 return(which(is.na(x)))  
}  
  
# 누락값 가진 컬럼 찾기  
apply(wt, 2, na\_idx) # sunshine, windgustspeed에 NA 존재

## $tempmin  
## integer(0)  
##   
## $tempmax  
## integer(0)  
##   
## $rainfall  
## integer(0)  
##   
## $sunshine  
## [1] 222 301 349  
##   
## $windgustspeed  
## [1] 212 272  
##   
## $windspeed  
## integer(0)  
##   
## $humidity  
## integer(0)  
##   
## $pressure  
## integer(0)  
##   
## $cloud  
## integer(0)  
##   
## $raintomorrow  
## integer(0)  
##   
## $xgrp  
## integer(0)

# 누락값 가진 행 전체 몇 개?  
sum(is.na(wt)) # 5개

## [1] 5

# 누락값 대체하기  
wt$sunshine <-   
 ifelse(is.na(wt$sunshine),  
 round(mean(wt$sunshine, na.rm = T), 2),  
 wt$sunshine)  
  
wt$windgustspeed <-   
 ifelse(is.na(wt$windgustspeed),  
 round(mean(wt$windgustspeed, na.rm = T), 2),  
 wt$windgustspeed)  
  
### 2  
  
sorted <- sort(abs(cor(wt)[10,-10]), decreasing = T)[1:3]  
names(sorted) # "cloud" "sunshine" "pressure"

## [1] "cloud" "sunshine" "pressure"

### 3  
# 훈련, 테스트 데이터 세트 구분하기  
train <- wt[wt$xgrp == 1,]  
test <- wt[wt$xgrp == 0,]  
# 분석에 필요 없는 xgrp 변수 제거  
wt <- wt[,-11]  
train <- train[,-11]  
test <- test[,-11]  
  
# 로지스틱 회귀분석  
model.glm <- glm(raintomorrow ~ .,   
 data = train, family = binomial)  
summary(model.glm)

##   
## Call:  
## glm(formula = raintomorrow ~ ., family = binomial, data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0518 -0.5012 -0.2861 -0.1408 2.6733   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 96.217822 43.189557 2.228 0.02589 \*   
## tempmin -0.063163 0.074763 -0.845 0.39820   
## tempmax 0.115877 0.088662 1.307 0.19123   
## rainfall -0.008449 0.048071 -0.176 0.86049   
## sunshine -0.281623 0.102919 -2.736 0.00621 \*\*  
## windgustspeed 0.082348 0.026301 3.131 0.00174 \*\*  
## windspeed -0.085119 0.038177 -2.230 0.02578 \*   
## humidity 0.045332 0.022123 2.049 0.04045 \*   
## pressure -0.100558 0.041417 -2.428 0.01518 \*   
## cloud 0.099213 0.112349 0.883 0.37719   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 255.65 on 255 degrees of freedom  
## Residual deviance: 161.81 on 246 degrees of freedom  
## AIC: 181.81  
##   
## Number of Fisher Scoring iterations: 6

# 새로운 데이터 입력과 예측값 만들기  
newdata <- data.frame(tempmin=17.2, tempmax=25.8, rainfall=0, sunshine=8.8, windgustspeed=41, windspeed=6, humidity=74, pressure=1011.5, cloud=7)  
pred <- predict(model.glm, newdata = newdata)  
  
round(pred, 2) # 비가 올 확률은 0.84

## 1   
## 0.84

(raintomorrow <- ifelse(pred >= 0.5, 1, 0)) # 비가 올 것이다.

## 1   
## 1

### 4  
pred1 <- predict(model.glm, newdata = test)  
test$pred1 <- ifelse(pred1 >= 0.5, 1, 0)  
  
table(pred = test$pred1, real = test$raintomorrow)

## real  
## pred 0 1  
## 0 92 6  
## 1 3 9

pro <- prop.table(table(pred = test$pred1, real = test$raintomorrow))  
  
## 정확도 정답: 0.9181818  
accuracy <- pro[1,1] + pro[2,2]  
accuracy

## [1] 0.9181818