AIR Decision Tree HW "정확도 0.65 이상으로 올리기"

201811526 이은주

[코드]

#AIR DecisionTree HW "정확도 0.65 이상으로 올리기"

data <- read.csv("C:\\Users\\이은주\\Desktop\\AIR 동아리\\0430과제\\credit.csv")

str(data) #전반적인 자료를 나타낸다.

install.packages("C50")

install.packages("gmodels")

install.packages("randomForest")

library(C50)

library(gmodels) #CrossTable()명령어를 실행하기 위한 패키지

library(randomForest)

table(data$default) #default는 채무불이행 자료 -> 1 또는 2로 표현되어있음.

data$default <- ifelse(data$default==1, "no", "yes")

#default를 타겟으로 하여 1또는2로 표현된 자료를 factor 자료로 표현하기 위해

#default 데이터가 1이면 'no', 1이 아닌 2는 'yes'라고 표현한다.

data$default <- as.factor(data$default) #default자료를 factor로 바꾼다.

str(data$default)

table(data$checking\_balance)

table(data$savings\_balance)

summary(data$months\_loan\_duration)

summary(data$amount)

##랜덤하게 있는 객체들이 아니기때문에 난수로 추출

set.seed(2018)

train\_sample <- sample(1000,900) #1-1000 중 900개의 숫자 추출

data\_train <- data[train\_sample,] #추출된 숫자(900개)에 대한 데이터 자리수

#그 자리에 대한 데이터 = data\_train

data\_test <- data[-train\_sample,] #추출하지 않은 나머지 100개의 자료 = data\_test

prop.table(table(data\_train$default)) #prop.table은 비율을 나타낸다.

prop.table(table(data\_test$default))

#채무불이행이 약 30%정도로 잘 나눠진걸 확인

data\_options <- C5.0Control(winnow=T, noGlobalPruning=F, CF=0.5)

#C5.0의 함수 파라미터를 생성하는 함수

#winnow -> 입력 필드에 대해서 사전에 필드가 유용한지 측정한 다음 유용하지 않는 경우 배제하고 모델링

#noGlobalPruning -> 전역적 가지치기 여부를 결정(전체적으로 만들어진 Tree구조에서

# 가지치기를 수행하는데 강도가 약한 sub-tree자체를 삭제

#CF -> 지역적 가지치기의 강도를 조정(값이 작을수록 가지치기 정도가 강해짐)

data\_model <- C5.0(data\_train[-17], data\_train$default, control = data\_options, trials=5)

#train데이터에서 타겟인 default(17번쨰변수)를 제외한 데이터셋을 넣고,

#타겟인 default를 지정해서 모델생성

#data\_options에 따라 C5.0 함수 사용

#trials -> 여러개의 분류모형에 의한 결과를 종합하여 분류의 정확도를 높인다. -> 부스팅 반복횟수를 지정

summary(data\_model) ##모델의 정보확인가능

#변수의 중요도 확인

da\_rf <- randomForest(default~., data\_train, importance=T)

varImpPlot(da\_rf)

#가장 낮은 변수 제거하여 모델생성

data\_model\_2 <- C5.0(data\_train[-c(17,20)], data\_train$default, control = data\_options, trials=5)

data\_pred <- predict(data\_model\_2, data\_test)

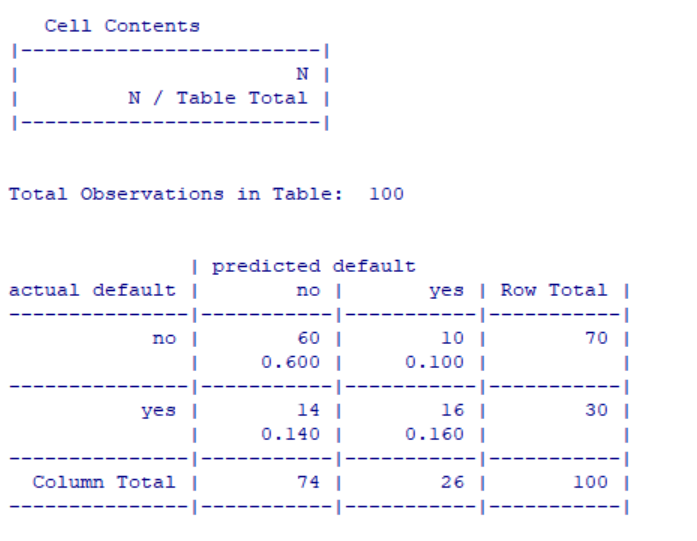
##data\_model\_2 모델에 data\_test를 적용시킨 예측값들 출력

CrossTable(data\_test$default, data\_pred,

prop.chisq=FALSE, prop.r=FALSE, prop.c=FALSE,

dnn=c('actual default','predicted default'))

[결과]



* 정확도가 76%로 증가되었음을 알 수 있습니다.