(B)

2018123223 경영학과 강지훈

# 1. 데이터 불러오기

Csv 파일 형태로 된 WA\_Fn-UseC\_-Telco-Customer-Churn.csv를 불러옵니다. 단 이때 캐릭터 변수를 factor로 불러 옵니다.

# 2. 데이터 ID 열 제거

분석에서 큰 도움이 되지 않는 ID 열을 전부 제거합니다

# 3. 결측치 확인과 제거

데이터 분석에서 왜곡된 해석을 야기할 수 있는 결측치를 확인하고 제거합니다

# 4. 데이터 factor

분석을 진행하기 위해서 factor화를 진행합니다.

# 5. Binning을 위한 점검을 진행합니다

박스 플롯을 통해 binning 방식을 확인합니다.

# 6. Binning 후 factor화와 split data

5번을 통해서 각각 10 / 10 / 1000 단위로 나눠 binning하기로 결정하고 그대로 factor화 합니다.

0.6 : 0.4로 데이터를 나눕니다

# 7. Navie Bayes 적용

지금까지 전처리 된 데이터를 나이브 베이스에 적용합니다

# 8. Decision tree 적용

# 8.1 binning 된 기존 데이터 활용

지금까지 전처리 한 데이터를 Decision tree를 적용 합니다

# 8.2 binning 전 Continuous 데이터 활용

# 8.2.1 데이터 전처리 및 split

# 8.2.2 Decision Tree 적용

Binning 하지 않고 factor화 되지 않은 원래 데이터 그 자체로 남아있는 Continuous 변수를 활용하여 Decision tree를 적용 합니다.

# 9.1 기존 데이터에 randomForest적용

# 1~ 6까지 전처리한 데이터를 이용해서 randomForest를 적용합니다.

# 9.2 binning 전 Continuous 데이터 활용

Binning 하지 않고 factor화 되지 않은 원래 데이터 그 자체로 남아있는 Continuous 변수를 활용하여 RandomForest를 적용 합니다.

# 10. NN을 위한 변수 7개 선택

varImpPlot(rf)을 이용해 7개 변수를 찾아내도록 합니다.

Contract, tenure, MonthlyCharges, InternetService, OnlineSecurity TotalCharges, TechSupprot이 채택됩니다/

# 11. MeanDecreaseAccuracy를 사용한 neuralnet을 위한 데이터 및 nnet 불러오기

뉴럴넷 작업을 위한 데이터와 library를 불러옵니다.

# 12. 변수 선택, 결측치 제거

#10에서 선택된 변수를 채택하고 결측치를 제거합니다

# 13. 정규화

factor화를 바로 할 수 없는 변수들의 정규화를 진행합니다

# 14. cbind와 paste를 이용한 더미화

더미화 할 수 있는 변수들을 더미화 해서 진행합니다.

# 15. 더미화 된 변수 중 기존 변수 및 Churn 제거

더미화 된 변수들을 남겨 놓고 더미화에 사용된 변수와 Churn을 제거합니다

# 16. split data

데이터를 6:4로 분할합니다

# 17. neuralnet 적용 및 정확도 확인

뉴럴넷을 적용하고 Confusionmatrix를 확인합니다.

# 보너스: Accuracy 높이기 위한 전략

RamdomForest의 오버피팅 이슈를 피하기 위해 Ntree값을 조정해 적용해 봅니다. Ntree는 400을 활용해봅니다.