```
##### Chapter2 데이터 전처리와 모형 평가 #####
library("caret")

### 2.1. 서론
# 분석 과정에서 고려해야 할 데이터의 전처리 과정과 구축된 모형에 대한 평가
# 데이터의 전처리는 모형을 구축하기 이전단계에서 수행
# 데이터의 전처리는 데이터의 정제, 정규화, 변환 및 변수 추출과 선택 등을 포함
# 또한, 구축된 모형에 대한 평가방법과 평가를 위한 여러 가지 척도를 소개한다.



### 2.2 데이터 전처리
# 이 절에서는 {caret} 패키지를 이용하여 데이터 전처리와 관련된 다음의 주제를 다룬다.
# 영- 과 영근처 분산 예측변수의 처리
# 상관된 예측변수 식별: 중복변수 제거
# 예측변수의 변환
# 기타 전처리 방법

# {caret} 패키지는 예측변수를 전처리하는 몇 가지 함수를 제공
# {caret} 패키지는 모든 데이터를 수치형으로 가정한다.
# 즉, 요인은 model.matrix{stats}, dummyVars{caret} 함수 또는 다른 방법을 통해 더미 변수로 변환되었다고 가정



### 2.2.1 영-과 영근처- 분산 예측변수의 처리
# 일부 상황에서 한 개의 값만을 취하는(영-분산) 예측변수를 생성할 수 있다.
# 트리 기반 모델 제외한 많은 모형에서, 영-분산 예측변수는 모형을 망가뜨리거나 불안정한 적합의 원인이 된다.
# 마찬가지로 예측변수는 매우 낮은 빈도로 발생하는 몇 개의 값만을 취할 수도 있다.
# 여기서 주의할 점은 이 예측변수들이 분할이 될 때 영-분산 예측변수가 되거나,
# 일부 샘플이 모형에 과도한 영향을 미치게 되는 경우이다.
# 따라서 영근처-분산 예측변수는 모형화 이전에 식별되고 제거되어야 한다.

# mdrr{caret} 자료에서 nR11 변수는 매우 불균형적으로 몇 개의 수치만을 취한다.
data(mdrr)
data.frame(table(mdrrDescr$nR11)) # 3개의 값만 가지는 데이터

# 영근처-분산 예측변수를 식별하기 위해 다음 두 개의 측도를 소개한다.
# freqRatio : 빈도비율, (일순위 빈도값) / (이순위 빈도값), 정상 데이터에선 1에 가깝고, 비정상 데이터에선 매우 큰 값을 가진다.
# percentUnique : 유일값들의 비율, (값의 종류) / (전체 표본의 수) * 100, 데이터의 집중도가 높아질수록 0에 가까워진다.
# 빈도비율이 임계 값보다 크고, 유일 값들의 비율이 임계값 보다 작으면 예측변수가 영-분산에 가깝다고 간주할 수 있다.
# 이산 균일 분포와 같이 고르게 분포된 데이터를 잘못 판단하지 않기 위해서 두 가지의 기준을 사용하는 것이 바람직하다.

# nearZeroVar함수
# 빈도비율, 유일값들의 비율 그리고 각변수들의 영분산, 영근처분산의 유무를 알수 있다.
# 디폴트 값, 빈도비율이 19크고, 유일값의 비율 10%보다 작으면 영근처분산으로 분류
# 옵션 : saveMetrics = TRUE -> 아래의 4가지 수치를 알려주며, 자세하게 출력
# freqRatio : 빈도비율, (일순위 빈도값) / (이순위 빈도값), 정상 데이터에선 1에 가깝고, 비정상 데이터에선 매우 큰 값을 가진다.
# percentUnique : 유일값의 비율, (값의 종류) / (전체 표본의 수) * 100, 데이터의 집중도가 높아질수록 0에 가까워진다.
# zeroVar : 영 분산
# nzv : 영 근처분산
nzv <- nearZeroVar(mdrrDescr, saveMetrics = TRUE)

# 데이터 요약 및 상위 데이터 출력
str(nzv) # 4가지 변수에 대한 342개 데이터
nzv[nzv$nzv, ][1:10,]

# 영근처 분산을 가지는 경우
dim(mdrrDescr) # 영근처 분산을 제거하기 전에 요약, 총데이터(528) 데이터(342)
nzv <- nearZeroVar(mdrrDescr) # 영근처 분산 데이터의 자세한 수치가 아닌 위치를 반환
```

```
head(nzv) # 영근처 분산 45개의 위치 중 앞에 6개만 출력
filteredDescr <- mdrrDescr[, -nzv] # 영근처 분산을 제외하는 과정
dim(filteredDescr) # 영근처 분산을 제거한 후에 요약, 총데이터(528) 데이터(297), 45개가 제거되었다.
```

```
### 2.2.2 상관된 예측변수의 식별 : 중복변수 제거
# 상관관계가 있는 예측변수에 대해서도 잘 작동하는 일부 모형이 있지만(예: PLS 회귀), 다른 모형들은 예측변수들 간의 수준을 줄이는 것이 좋다.
# findCorrelation() 함수는 제거해야 할 예측변수를 제공해준다.

# 중복변수를 제거하기 전 데이터 요약
descrCor <- cor(filteredDescr) # 영근처 분산을 제외한 변수들 간의 상관계수 구하기
summary(descrCor[upper.tri(descrCor)]) # 상관계수 처리를 하기 전 요약, min과 max를 보아 상관계수가 1인 관계가 있음을 인지

# 높은 상관계수 제거
highCorr <- sum(abs(descrCor[upper.tri(descrCor)]) > .999) # 상관계수가 0.999이상인 변수들의 개수
highlyCorDescr <- findCorrelation(descrCor, cutoff = 0.75) # 상관계수가 0.75보다 높은 값들의 위치 반환
filteredDescr <- filteredDescr[, -highlyCorDescr] # 상관계수가 0.75가 넘어가는 값들을 제거

# 중복변수를 제거한 후 데이터 요약
descrCor2 <- cor(filteredDescr) # 수정된 데이터의 상관계수
summary(descrCor2[upper.tri(descrCor2)]) # 상관계수 처리를 한 후 요약, min, max가 0.75보다 작은 것을 볼 수 있다.
```

```
### 2.2.3 예측변수의 변환

## 중심화와 척도화
# preProcess() 함수는 중심화와 척도화를 포함하여 예측변수에 대해 많은 연산을 제공한다. 실제로 데이터를 전처리하지 않는다.
# preProcess() 함수는 측정 데이터 셋(훈련용 자료)으로부터 요구하는 것을 추정한 다음, 이 값을 재계산하지 않고 임의의 데이터 셋에 적용한다.
# preProcess() 함수는 각 연산에 필요한 모수를 제공한다.
# predict.preProcess() 함수는 특정 데이터 셋에 이를 적용하는 데 사용된다. 훈련용 셋과 검증용 셋을 전처리 하는데 사용된다.

# 영근처 분산과 상관계수 처리를 한 데이터에 대해 훈련용 셋과 검증용 셋을 생성
set.seed(200)
inTrain <- sample(seq(along = mdrrClass), length(mdrrClass)/2) # 두 개의 데이터 셋으로 나누기 위해 기준 생성
training <- filteredDescr[inTrain, ] # 훈련용 데이터 셋 생성
test <- filteredDescr[-inTrain, ] # 검증용 데이처 셋 생성

# 중심화와 척도화 진행
# method option = "center"(중심화), "scale"(척도화), "BoxCox"(박스콕스변환), "ranges"(-0 과 1사이의 값으로 데이터를 변환)
preProcValues <- preProcess(training, method = c("center", "scale")) # 중심화와 척도화에 필요한 모수 추정
trainTransformed <- predict(preProcValues, training) # 훈련용 셋에 적용(중심화와 척도화 전처리 진행)
testTransformed <- predict(preProcValues, test) # 검증용 셋에 적용(중심화와 척도화 전처리 진행)

## box-cox 변환
preProcValues2 <- preProcess(training, method = "BoxCox") # 박스콕스 변환에 필요한 모수 추정
trainBC <- predict(preProcValues2, training) # 훈련용 셋에 적용(박스콕스 전처리 진행)
testBC <- predict(preProcValues2, test) # 검증용 셋에 적용(박스콕스 전처리 진행)
```

```
### 2.2.4 기타 전처리 방법
# 번주형 자료의 처리를 위한 더미변수의 생성, 자료의 열들간의 선형종속성 관계 파악, 결측값 대치,
# 새로운 예측변수 생성을 위한 군집거리 계산 등을 소개한다.

## 더미변수 생성
# {caret} 피키지에 dummyVars() 함수는 하나 이상의 요인으로부터 완전한 더미변수 집합을 생성해준다.
# 다음의 예제에서는 두 가지의 방법으로 더미 변수를 만들기 위한 모수를 만들고, predict() 함수를 적용하여 더미 변수를 만든다.
```

```
# etitanic{earth} 자료는 두 개의 요소형 변수 pclass(1st, 2cd, 3rd)와 sex(famale, male)를 포함한다.
# install.packages("earth")
library(earth) # 패키지 다운
data(etitanic) # 데이터 불러오기
str(etitanic) # 데이터 구조 확인, 6개의 변수(pclass, survived, sex, age, sibsp, parch), 1046개 데이터

# dummyVar{caret} 함수 사용 : 절편 없음 -> lm()을 비롯한 일부 모형에 유용하지 않을 수도 있다.
dummy.1 <- dummyVars(survived ~ ., data = etitanic) # 더미변수에 필요한 모수 추정
head(predict(dummy.1, newdata = etitanic)) # predict() 함수를 통해 더미 변수를 만든다.

# model.matrix{stats} 함수 사용 : 절편 있음
head(model.matrix(survived ~ ., data = etitanic)) # 더미변수에 필요한 모수 추정과 더미변수 집합 생성을 동시에 진행



## 선형종속성
# {caret} 패키지의 findLinearCombos() 함수를 사용하여 선형종속성을 제거한다.
# 행렬의 Q,R 분해를 사용하여 선형결합의 집합을 열거해주고 선형종속성을 없애기 위해 제거 되야할 열 위치를 제공

# 선형독립성이 존재하지 않은 데이터, 즉 종속성이 있는 데이터 선언(예, 2열 + 3열 = 1열)
ltfrDesign <- matrix(0, nrow = 6, ncol = 6)
ltfrDesign[, 1] <- c(1, 1, 1, 1, 1, 1)
ltfrDesign[, 2] <- c(1, 1, 1, 0, 0, 0)
ltfrDesign[, 3] <- c(0, 0, 0, 1, 1, 1)
ltfrDesign[, 4] <- c(1, 0, 0, 1, 0, 0)
ltfrDesign[, 5] <- c(0, 1, 0, 0, 1, 0)
ltfrDesign[, 6] <- c(0, 0, 1, 0, 0, 1)

# 선형종속성을 위반하는 열의 위치 반환
comboInfo <- findLinearCombos(ltfrDesign) # 선형종속성을 없애기 위해 제거할 열의 위치 벡터를 반환
comboInfo$linearCombos[[1]] # 첫번재 선형종속성 그룹(3 1 2)
comboInfo$linearCombos[[2]] # 두번째 선형종속성 그룹(6 1 4 5)
comboInfo$remove # 제거되야할 열의 위치, (3 6)

# 선형종속성 처리
ltfrDesign[, -comboInfo$remove]



## 결측값 대치
# preProcess() 함수는 훈련용 자료에서의 정보를 가지고 데이터 셋의 결측값을 대치하는데 사용할 수 있다. 값을 리턴하진 않는다.
# 특정 데이터 셋에 이를 적용하여 값을 확인하기 위해서 RANN 패키지의 predict()함수를 사용한다.
# k-근접 방법 : 임의의 하나의 표본에 대해, k개의 가장 가까운 이웃을 훈련용 자료에서 발견하고, 이들 값의 평균등을 구하여 대치한다.
# k-근접 방법 : 이 방법을 사용하면 모드 옵션이 무엇이든 관계없이 preProcess()가 중심화와 척도화를 수행하게 해준다.
# 배깅 트리 모형 방법 : 데이터의 각 예측변수에 대해, 훈련용 자료의 다른 모든 예측변수를 사용하여 배깅 트리가 만들어진다.
# 배깅 트리 모형 방법 : 새 데이터의 결측값을 배깅 트리을 사용할 수 있으며, 더 강력한 대치 방법이지만 기용이 훨씬 높다.

# airquality{caret} 데이터 불러오기
install.packages("RANN") # predict()함수를 불러오기 위해 패키지 선언
library(caret) # preProcess() 함수의 경로 확인
library(RANN) # predict() 함수의 경로 확인
data(airquality) # 데이터 불러오기

# 결측값을 대체하기 전 데이터 요약
summary(airquality) # 데이터 요약(결측값 개수 확인, Ozone 37개, Solar.R 7개)

# 결측값을 k-근접 방법으로 대체하는 과정
# method option = "center"(중심화), "scale"(척도화), "BoxCox"(박스콕스변환), "knnImpute"(k-근접 방법),
#                                                                  "ranges"(-0 과 1사이의 값으로 데이터를 변환,)
imp.1 <- preProcess(airquality, method=c("knnImpute")) # k-근접 방법에 필요한 모수 추정
imp.2 <- predict(imp.1, airquality) # 추정된 모수를 사용하여 k-근접 방법 대치 실행
```

```
# 결측값을 대체한 후 데이터 요약
summary(imp.2) # 데이터 요약(결측값이 사라졌음)


## 군집거리 계산
# {caret} 패키지에는 군집 중심까지의 거리를 기반으로 새로운 예측변수를 생성하는 함수가 포함되어 있다.
# 요인 변수의 각 수준에 대해 군집의 중심과 공분산 행렬이 계산된다.
# 새로운 표본에 대해, 각 군집중심까지의 마할라노비스 거리가 계산되고 이 값은 추가 예측변수로 사용될 수 있다.
# lassDist() 함수는 표본보다 군집 내에 예측변수가 더 많은 경우, pca= 와 keep 옵션을 통해 특이 공분산행렬 문제를 해결한다.
# predict.classDist() 함수는 군집거리를 생성하는 데 사용된다. 디폴트로 거리가 기록되고, trans= 옵션을 통해 변경할 수 있다.

# iris 데이터에서 랜덤하게 train(훈련용 셋)집단 선정
trainSet <- sample(1:150, 100)
distData <- classDist(iris[trainSet, 1:4], iris[trainSet, 5]) # classDist(데이터 셋, 분류기준)
distData$values # 훈련용 자료로부터 군집 중심과 공분산 행렬 계산

# test집단으로 군집 중심까지의 마할라노비스 거리 계산
newDist <- predict(distData, iris[-trainSet, 1:4]) # iris[-trainSet, 1:4] : 검증용 셋, 50개의 데이터
newDist # 거리가 가까울수록 그 종류일 확률이 높다. 예 1번은 setosa일 확률이 높다.

# test집단에 대한 군집거리의 산점도 행렬
# 각 집단이 잘 구분되어 있으면 특징이 뚜렷한 것이다.
# “Setosa는 versicolor, virginica와 구분이 잘 되지만, versicolor와 virginica는 비슷한 면이 있어 구분이 잘 안 됨.”이라고 해석 가능
splom(newDist, groups = iris$Species[-trainSet], auto.key=list(columns=3))



### 2.3 모형평가

### 2.3.1 최적의 부분 회귀모형의 선택 기준
# 최적의 부분 회귀모형의 선택은 예측변수들의 가능한 모든 부분집합을 예측변수로 하는 회귀모형을 적합하고,
# 이 가운데 아래의 기준에 잘 부합하는 모형을 찾는 방법이다.
```

결정계수 $(R^2) = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

수정결정계수 $(R_a^2) = 1 - (\frac{n-1}{n-p})\frac{SSE}{SST}$

평균제곱오차 $(MSE) = \frac{SSE}{n-p} = \frac{\sum(y_i - \hat{y}_i)^2}{n-p}$

$Mallow's\ C_p\ (C_p) = p + \frac{(MSE - \hat{\sigma^2})(n-p)}{\hat{\sigma^2}} = \frac{SSE}{\hat{\sigma^2}} + 2p - n$

$\therefore p =$ 모수의 수, $\hat{\sigma^2} =$ 모든 예측변수를 포함한 적합모형의 평균제곱오차

$\therefore SSE = p$개의 예측 변수로 적합한 모형의 오차제곱합

```
# 위에 기준에 따른 변수 선택 절차는 다음과 같다.
# 1) 결정계수는 p의 증가에 따라 증가함수이다. 따라서 증가가 둔화되는 시점의 p를 선택한다.
# 2) 설명력이 떨어지는 예측변수가 추가되어도 값이 증가하는 단점을 보완한 수정된결정계수의 결과와 동일하다.
# 3) MSE가 최소가 되는 p를 선택한다. (수정된 결정계수의 결과와 동일하다.)
# 4) Cp의 값이 p와 가장 가까운 값을 가지는 p를 선택한다.



### 2.3.2 정보 기준과 PRESS
## 정보기준
# Cp가 실제보다 모형 간에 더 큰 차이가 있는 것처럼 보이게 하는 경향 때문에 일부 전문가들은 정보 기준이 Cp보다 더 현실적인 방법이라고 생각
# 세 종류의 정보기준은 모두 작은 값을 가질수록 우수하다라고 할 수 있다.
# BIC 는 AIC에 배해 모수의 수에 더 큰 벌점을 부여하므로 좀 더 단순한 모형을 선호하게 된다.
# 즉, 이는 과대 적합의 영향이 있다고 파악되는 AIC의 대한 보완으로 BIC가 나왔다고 볼 수 있다.
```

$Akaike's$ 정보기준 : $AIC = n\ln(\frac{SSE}{n}) + 2p$

$Bayesuan$ 정보기준 : $BIC = n\ln(\frac{SSE}{n}) + p\ln(n)$

$Amemiya's$ 예측기준 : $APC = \frac{n+p}{n(n-p)}SSE$

$\therefore n$ = 표본의 크기이며, p와 SSE의 정의는 위와 같다.

예측 제곱합(PRESS)
데이터 셋을 둘로 나누지 않고 모형의 예측력을 통해 평가하는 방법, 값이 작을수록 예측력이 우수하다고 할 수 있다.

$PRESS = \sum_{i=1}^{n}(y_i - \hat{y}_{i(i)})^2$, $\hat{y}_{i(i)}$는 i번째 자료를 제외하고 적합한 모형으로부터 i번째 값을 추정한 것이다.

예측 결정 계수(R^2_{pred})
이 값은 PRESS 보다 더 직관적으로 해석 될 수 있으며, 데이터 셋을 둘로 나누지 않고 예측력을 비교할 수 있어 유용하다.

$R^2_{pred} = 1 - \frac{PRESS}{SST}$, (이 값이 0보다 작을 때는 영으로 간주함)

PRESS와 예측결정계수는 모형 추정에 포함되지 않은 자료를 이용하여 계산되므로 과적합을 방지하는 데 도움이 된다.
과적합은 모형 적합에 사용된 데이터에 대해서는 우수한 적합을 제공하지만, 새로운 관측값에 대해서는 유용한 적합을 보이지 못하는 것을 의미

교차 타당법
데이터 셋을 훈련용 셋(모형구축에 사용될)과 평가용 셋(예측력 평가에 사용될)으로 나누어 모형을 평가하는 방법으로,
데이터 양이 충분히 많은 경우에는 두 데이터 셋의 비율을 50:50으로 랜덤하게 나누어 적용한다.

K-중첩 교차타당법
데이터 양이 충분하지 않는 경우 K 조각으로 나누어(K-1)조각으로 모형 구축한 뒤 1조각으로 예측을 수행하는 방법
이 절차를 K번 반복한다. 각 조각에 대한 제곱예측오차를 더하여 교차타당법의 측도로 이용한다.(조각은 데이터 셋이라고 생각하면 편함.)

LOO 교차 타당법
K-중첩 교차 타당법에서 K=n인 경우에 해당, 즉 한 개를 제외하고 모형을 구축한뒤 남은 한 개를 추정하는 과정을 반복.
예측 오차의 추정치는 PRESS와 동일하다.

2.3.4 데이터 마이닝에서의 모형평가
예측 모형에 대한 평가는 보통 훈련용 자료에 의해 구축된 모형을 검증용 자료에 적용하여 평가한다,
모형 평가에 사용되는 측도로 다음과 같이 적용한다.

범주형 반응변수에 대해서는 정오분류표에 기반한 (정분류율, 민감도, 특이도) 등이 사용
연속형 반응변수에 대해서는 평균 절대 오차, 평균 제곱 오차 등이 사용된다.

이진 반응 변수의 경우
정오분류표는 예측결과가 두 개의 집단(C_1, C_2)으로 주어지는 경우 다음과 같이 정의된다.(전체 자료의 개수는 n)

		예측집단	
		C_1	C_2
실제집단	C_1	$f_{11}(T-P)$	$f_{12}(T-N)$
	C_2	$f_{21}(F-P)$	$f_{22}(P-N)$

정분류율or 정확도 : $\frac{f_{11}+f_{22}}{n}$

민감도(참일 것을 참으로 제대로 분류한 비율) : $\frac{f_{11}}{f_{11}+f_{12}}$

특이도(실제 거짓인 것을 거짓으로 제대로 분류한 비율) : $\frac{f_{22}}{f_{21}+f_{22}}$

정분류율의 민감도와 특이도의 가중합으로 표현될 수 있다.

정분류율 $= \frac{f_{11}+f_{12}}{n}$*(민감도) + $\frac{f_{21}+f_{22}}{n}$*(특이도)

연속형 반응변수의 경우

연속형 예측값에 대해 적용되는 평균절대오차, 평균제곱오차, 평균절대백분위오차는 다음과 같이 정의된다.

실제값은 y_i, 예측값은 $\hat{y}_i$라고 하자.

(평균 절대오차) $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$

(평균 제곱오차) $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

(평균 절대백분위오차) $MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \times 100$

모형 선택을 위한 비교방법

모형 선택을 위한 여러 가지 예측 모형 간의 비교 방법은 다음의 2가지가 있다.

신뢰구간(또는 검정)을 이용하는 방법 : 두 개의 예측모형 간의 비교를 할 때 사용이 된다.

ROC 곡선을 이용하는 방법 : 연속형 값으로 주어질 때 유용하다. ROC 곡선을 그리는 것은 PASS

ROC 곡선 아래쪽 면적이 클수록 모형의 성능이 평균적으로 우수함을 나타낸다.