5.4 새로운 데이터에 적용된 거래 데이터 모델에 대한 십분위 향상차트인 <그림5.12>에 대해서 다음을 답하시오.

d. 이 상황에서 모델의 성능에 대한 이 두가지 측도(오차율과 향상도)의 유용성에 대해 의견을 제시하시오.

>> 모든 것을 비사기라고 분류하면 오차율이 감소(8.65%->5.96%)한다. 하지만 모든 것을 비사기라고 분류함으로써 향상도는 0이 된다.(사기인 Class를 사기라고 분류하는 경우의 수는 0 : $n_{1,1}$=0). 그러므로 모형에서 둘의 관계를 파악하고 필요에 따라 적절한 오차율과 향상도를 가지는 기준값 설정이 필요하다.

5.7 <표 5.7>은 실제 값과 경향 모두를 보여주는 분류모델의 예측모델 검증결과의 일부분을 보여주고 있다.

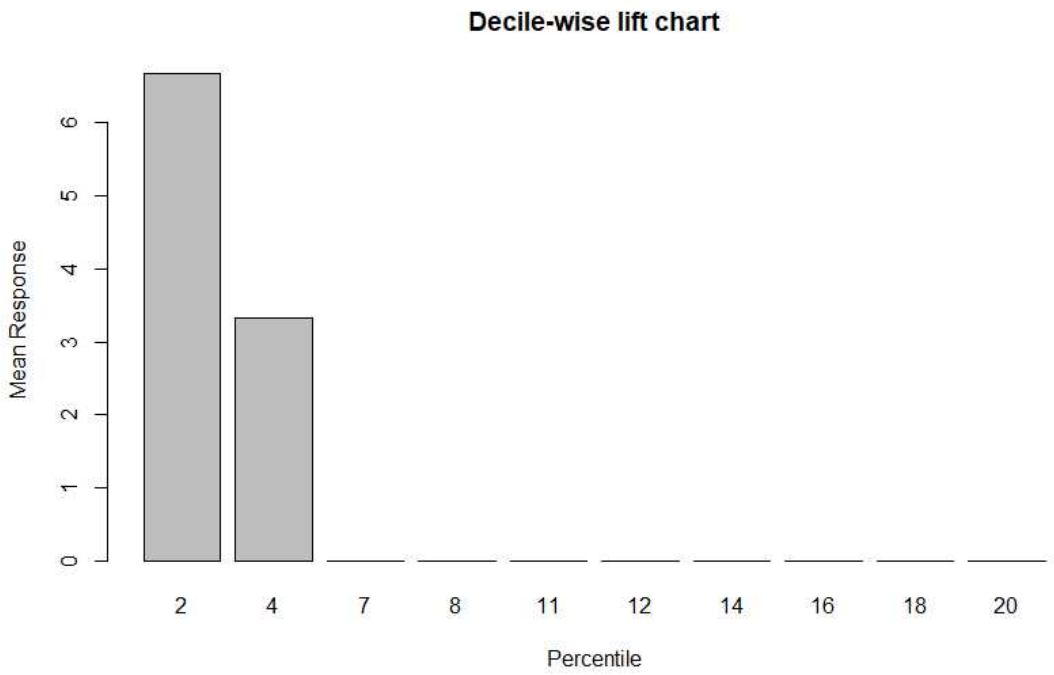
a. 0.25, 0.5, 0.75의 컷오프 값을 사용하여 오차율, 민감도, 특이도를 계산하시오.

```
setwd("c:/Rdata")
Homework.df <- read.csv("Homework.csv", header=TRUE)
library(caret)
library(e1071)

confusionMatrix(as.factor(ifelse(Homework.df$Probability > 0.25, "1", "0")), #예측값
                as.factor(Homework.df$actual), positive = "1") #실제값

confusionMatrix(as.factor(ifelse(Homework.df$Probability > 0.5, "1", "0")), #예측값
                as.factor(Homework.df$actual), positive = "1") #실제값

confusionMatrix(as.factor(ifelse(Homework.df$Probability > 0.75, "1", "0")), #예측값
                as.factor(Homework.df$actual), positive = "1") #실제값
```

컷오프값	오차율	민감도	특이도
0.25	0.4	1	0.5294
0.5	0.05	1	0.941
0.75	0.4	0.6667	1

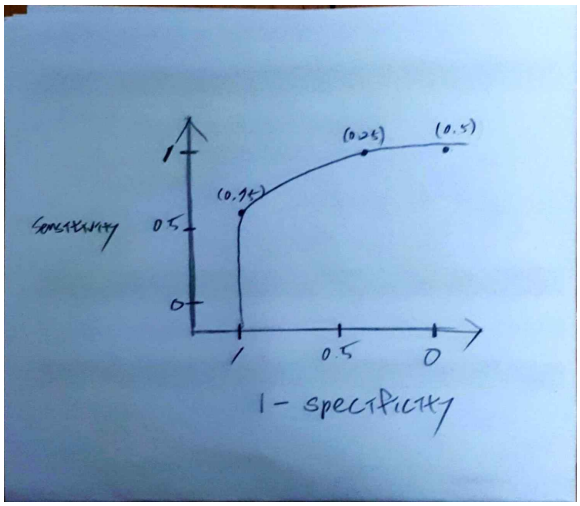
b. R에서 십분위 향상차트를 작성하시오.

```
library(gains)
gain <- gains(Homework.df$actual, Homework.df$Probability)
gain
barplot(gain$mean.resp/mean(Homework.df$actual), names.arg=gain$cume.obs)
```



> 각 분위마다 Y축(LIFT)값이 왜 나오는지 설명할 줄 알아야 함.
(10%:2/0.3=6.6, 20%:1/0.3=3.3)
> 컷오프의 값에 따라 십분위 차트를 그리면 틀림!

c. a에서 3개의 기준값을 이용해서 3쌍의(민감도, 특이도)를 이용하여 손으로 ROC곡선을 그려보시오.(무조건 시험문제 : 간단한 정오 분류표를 제시하고 ROC curve를 그려라!)



14.4 화장품 구매

a. 행렬의 여러 값들을 선택하여 그 의미를 설명하시오.

각 거래에서 어떤 품목의 화장품이 거래되었는지 확인할 수 있다. 행렬의 값 중에는 품목 두 개 구매를 의미하는 2가 있어, 나중에 데이터를 전처리하는 과정에서 count값을 incidence값으로 전환해야한다.

b. <표14.15>에 나타난 연관규칙 분석 결과를 보고, 다음에 답하시오.

(1) 첫 번째 행에서 '신뢰도' 출력의 의미와 어떻게 계산되는지 설명하시오.

첫 번째 행의 신뢰도가 0.3023255714라는 것은 조건부 항복집합{Blush, Concealer, Mascara, Eye.shadow, Lipstick}이 나왔을 때 조건부 항목집합{Blush, Concealer, Mascara, Eye.shadow, Lipstick}과 결론부 항목집합{Eyebrow.Pencils}이 포함된 거래의 수의 비율이다. 쉽게 말해 조건부 항목집합을 포함하는 거래의 약 30%가 조건부와 결론부를 동시에 포함하고 있다.

$$\text{신뢰도} = \frac{P(\text{조건부 항목집합 and 결론부 항목집합})}{P(\text{조건부 항목집합})}$$

(2) 첫 번째 행에서 '지지도' 출력의 의미와 어떻게 계산되는지 설명하시오.

첫 번째 행의 신뢰도가 0.013라는 것은 전체 거래의 수에서 조건부 항목집합{Blush, Concealer, Mascara, Eye.shadow, Lipstick}과 결론부 항목집합{Eyebrow.Pencils}이 동시에 포함되는 거래의 수의 비율이다. 쉽게 말해 전체 거래에서 조건부 항목과 결론부 항목의 집합 거래가 약 1.3%정도 빈출 된다고 말할 수 있다.

$$\text{지지도} = \frac{\text{조건부 항목집합과 결론부 항목집합을 동시에 포함하는 거래의 수}}{\text{전체 거래의 수}}$$

(3) 첫 번째 행에서 '향상'의 의미와 어떻게 계산되는지 설명하시오.

첫 번째 행에서 향상이 7.198228128 이라는 것은 해당 규칙이 높은 연관성을 가지고 있다는 뜻이다. 기준 신뢰도에 비해서 신뢰도가 높으면 의미있는 규칙이라고 판단할 수 있다.

$$\text{향상도} = \frac{\text{신뢰도}}{\text{기준 신뢰도}}$$

(4) 첫 번째 행에서 그 규칙이 표현하는 의미를 말로 설명하시오.

거래에서 Blush, Concealer, Mascara, Eye.Shadow, Lipstick을 사면 Eyebrow.Pencils을 살 가능성이 높다.

c. 이제 (Cosmetics.csv 파일에 저장된) 화장품 구매에 대한 전체 데이터세트를 사용하시오. R을 이용한 이 데이터에 연관규칙을 적용하시오. (기본 파라미터를 사용할 것)

(1) 출력된 처음 세 개의 규칙들을 말로 해석하시오.

```
Cos.df <- read.csv("Cosmetics.csv", header=TRUE)
incid.Cos.df <- ifelse(Cos.df > 0, 1, 0)
incid.Cos.mat <- as.matrix(incid.Cos.df[ , -1])
incid.Cos.trans <- as(incid.Cos.mat, "transactions")
rules <- apriori(incid.Cos.trans, parameter = list(target = "rules"))
inspect(head(sort(rules, by="lift"), n=3))
```

```
    lhs                              rhs           support confidence lift     count
[1] {Brushes}                     => {Nail.Polish} 0.149   1.0000000  3.571429 149
[2] {Blush,Concealer,Eye.shadow} => {Mascara}     0.119   0.9596774  2.688172 119
[3] {Blush,Eye.shadow}           => {Mascara}     0.169   0.9285714  2.601040 169
```

[1] Brushes를 사면 Nail.Polish를 사는 규칙의 지지도는 0.149, 신뢰도는 1, 향상도는 3.571이다. 즉 전체 데이터에서 빈출 정도는 14.9%에 해당하고, Brushes를 사면 무조건 Nai.Polish를 구매한다. 향상도는 1보다 크기 때문에 의미 있는 규칙이라고 말할 수 있다.

[2] Blush, Concealer, Eye.shadow를 사면 Mascara를 사는 규칙의 지지도는 0.119, 신뢰도는 0.9596774, 향상도는 3.571429이다. 즉 전체 데이터에서 빈출 정도는 11.9%에 해당하고. Blush, Concealer, Eye.shadow를 사면 약 95% 정도 Mascara를 구매한다. 향상도는 1보다 크기 때문에 의미 있는 규칙이라고 말할 수 있다.

[3] Blush, Eye.shadow를 사면 Mascara를 사는 규칙의 지지도는 0.169, 신뢰도는 0.9596774, 향상도는 2.601040이다. 즉, 전체 데이터에서 빈출 정도는 16.9%에 해당하고, Blush, Eye.shadow를 사면 약 92% 정도 Mascara를 구매한다. 향상도는 1보다 크기 때문에 의미있는 규칙이라고 말할 수 있다.