**[네 번째 과제]**

**IT응용시스템공학과**

**1494053**

**김희택**

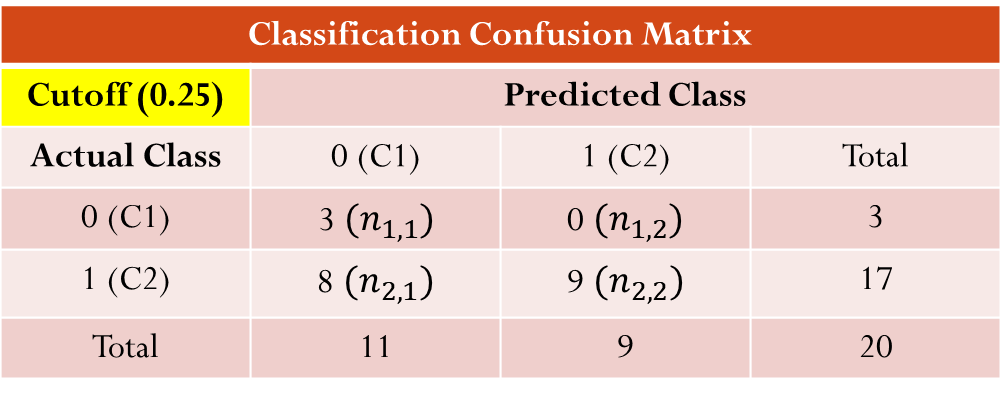
**5.4 새로운 데이터에 적용된 거래 데이터 모델에 대한 십분위 향상차트인 <그림 5.12>에 대해서 다음을 답하시오.**

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

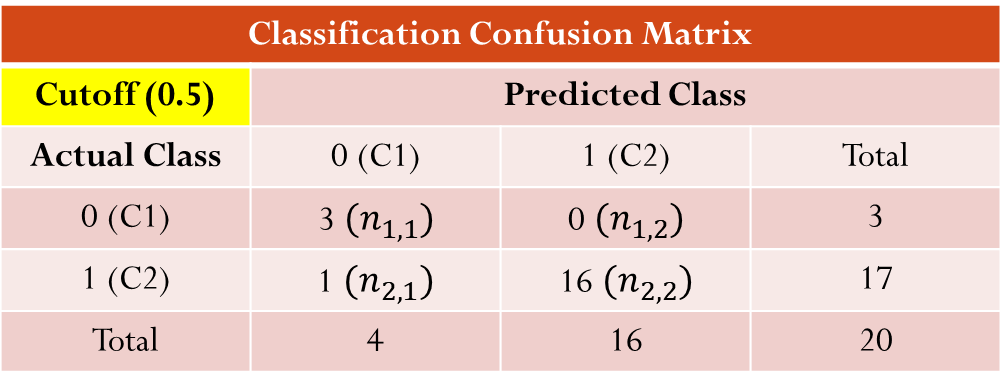
* 오차율이 초기보다 약 2.69% 감소하여 약 5.96% 라는 것은 모델의 성능이 더 정확해 졌다는 것을 의미한다. 또한 향상차트 정보는 누적예측 값이 높은 순서로 배열 되는 것을 한 눈에 확인할 수 있다. 즉 상위 20% (왼쪽에서 두 번째 막대)까지 누적예측 값이 가장 높은 것을 알 수 있다. 따라서 오차율이 줄어들고 향상도가 높은 것이, 랜덤선택 보다 더 정확한 데이터를 얻을 수 있으므로 모델의 성능이 향상된 것을 의미한다.

**5.7 <표5.7>은 실제 값과 경향 모두를 보여주는 분류모델의 예측모델 검증결과의 일부분을 보여주고 있다.**

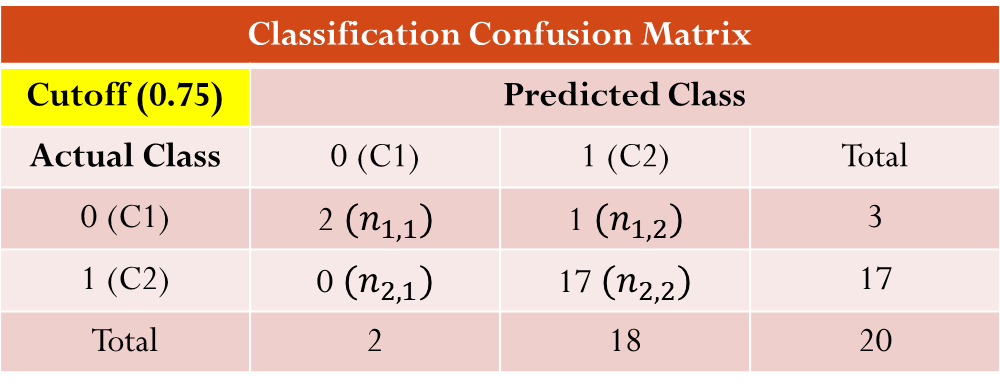
**a. 0.25, 0.5, 0.75의 컷오프 값을 사용하여 오차율, 민감도, 특이도를 계산하시오.**

****

* 오차율 = (잘못 분류된 레코드의 총합) / (전체 레코드) = 8 / 20 = 40%
* 만약 “C1”이 중요한 클래스라면
* 민감도 = “C1” 클래스가 정확하게 분류될 확률. 즉 = 3 / 3 = 100%
* 특이도 = “C2” 클래스가 정확하게 분류될 확률. 즉 = 9 / 17 = 약 53%

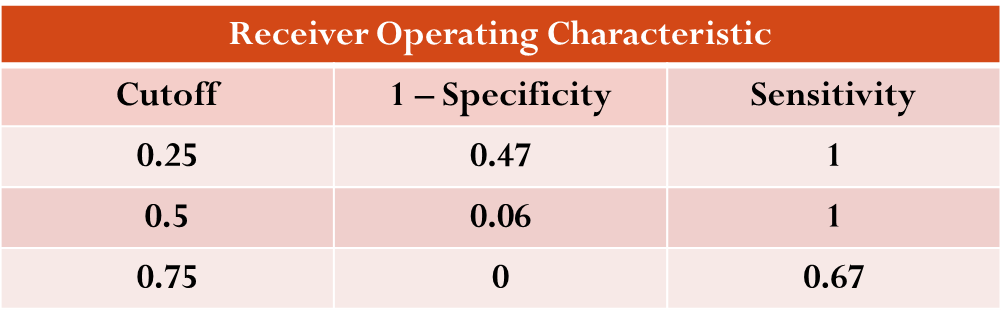
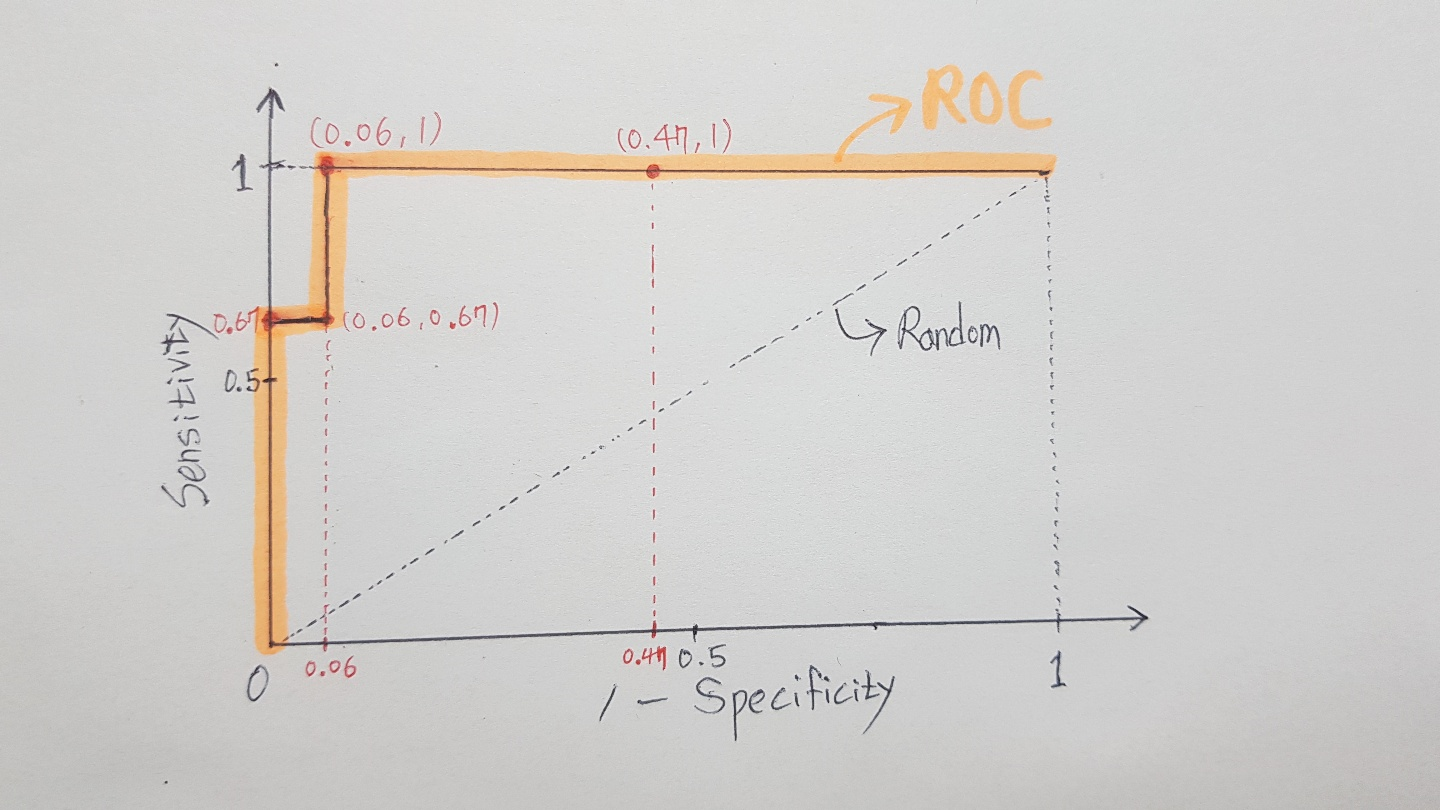
****

* 오차율 = (잘못 분류된 레코드의 총합) / (전체 레코드) = 1 / 20 = 5%
* 만약 “C1”이 중요한 클래스라면
* 민감도 = “C1” 클래스가 정확하게 분류될 확률. 즉 = 3 / 3 = 100%
* 특이도 = “C2” 클래스가 정확하게 분류될 확률. 즉 = 16 / 17 = 약 94%

****

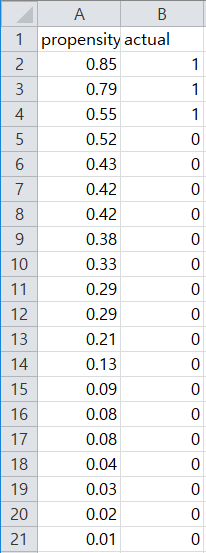
* 오차율 = (잘못 분류된 레코드의 총합) / (전체 레코드) = 1 / 20 = 5%
* 만약 “C1”이 중요한 클래스라면
* 민감도 = “C1” 클래스가 정확하게 분류될 확률. 즉 = 2 / 3 = 약 67%
* 특이도 = “C2” 클래스가 정확하게 분류될 확률. 즉 = 17 / 17 = 100%

**추가. a에서 3개의 기준값을 이용해서 구한 3쌍의 민감도, 특이도를 이용하여 손으로 ROC곡선을 그리시오.**

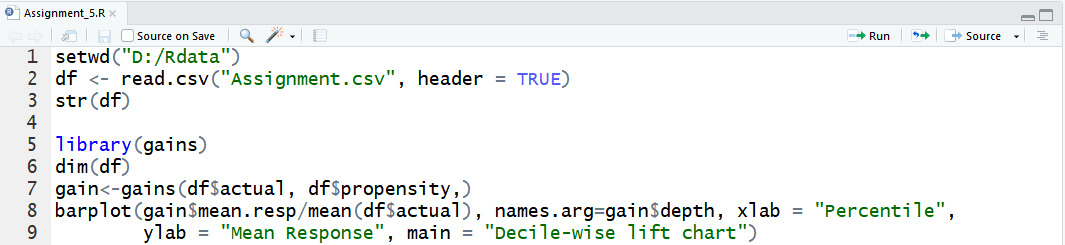
****

* 위에서 구한 민감도와 특이도를 통해 ‘1-Specificity’, ‘Sensitivity’를 도출한 후 ROC곡선을 그렸습니다.

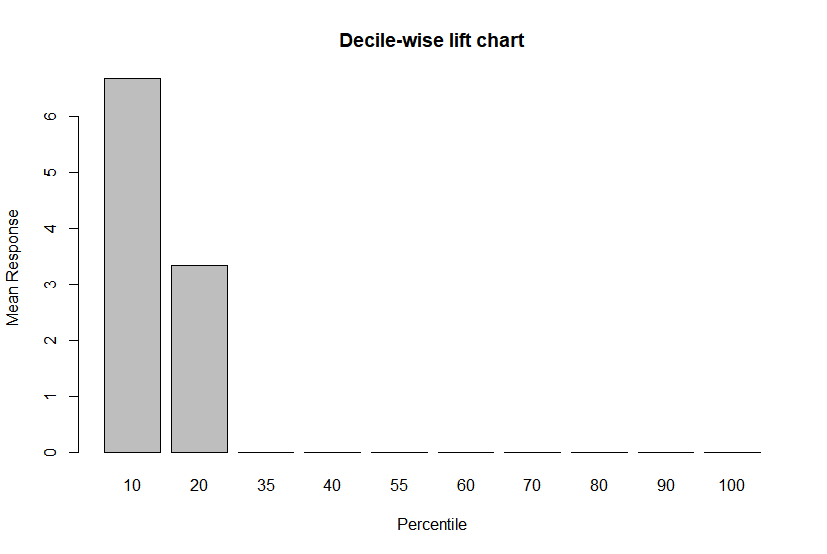
**b. R에서 십분위 향상차트를 작성하시오.**



* 먼저 주어진 <표 5.7>를 내림차순으로 정렬하여 csv 파일을 생성하였습니다.



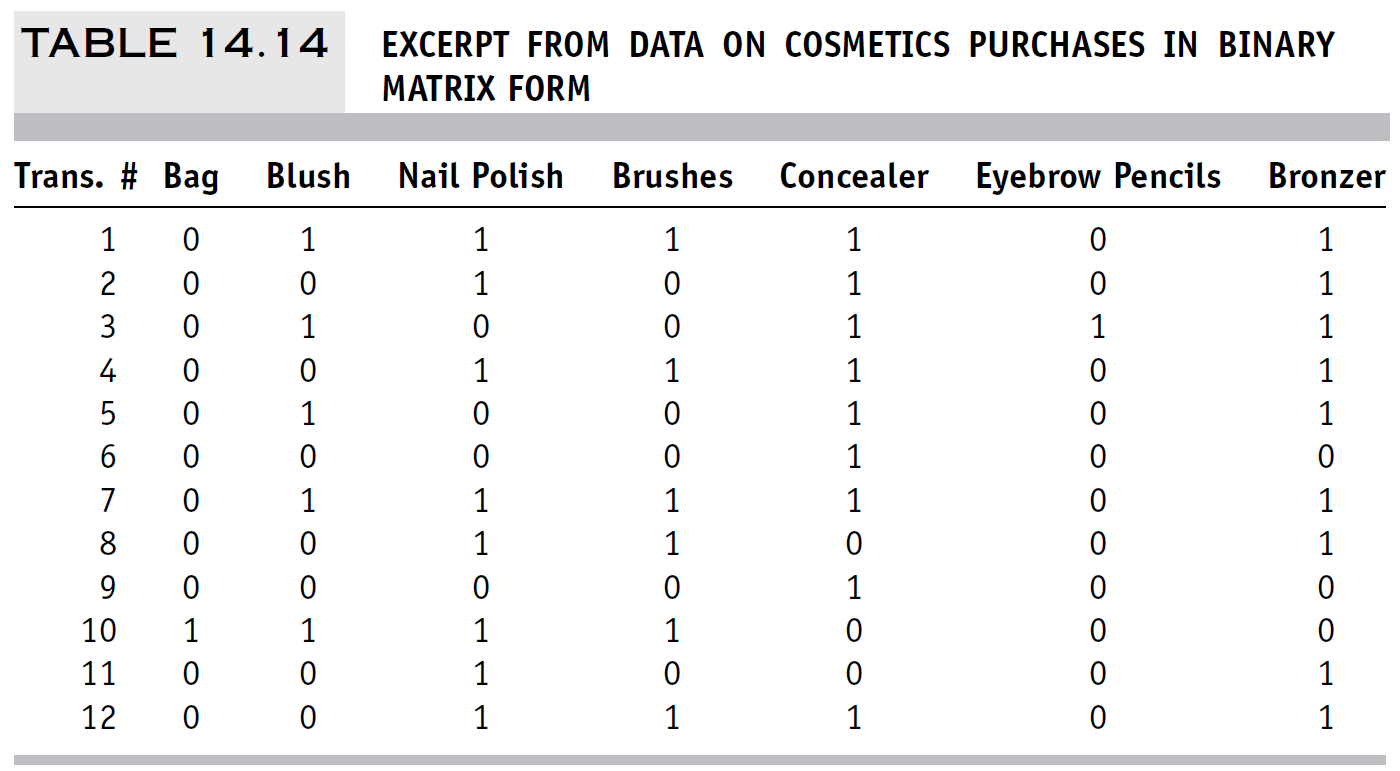
* 십분위 향상차트를 생성하는 R코드를 참고하였습니다. (교재 p.155)



* 십분위 향상차트 생성 결과입니다.

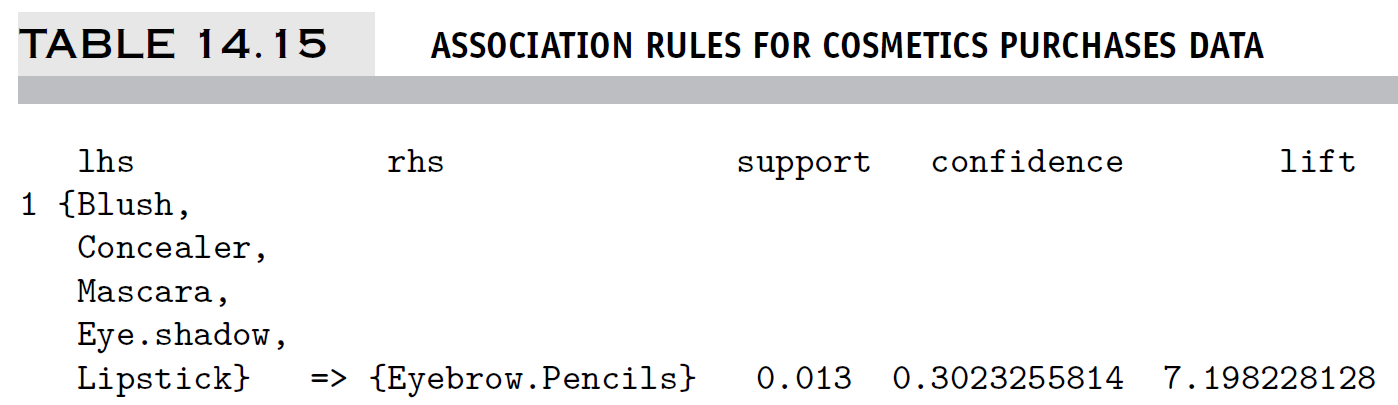
**14.4 화장품 구매 <표 14.14>와 같은 데이터와 <표 14.15>의 출력결과는 거대 체인 화장품점의 화장품 구매 데이터베이스(Cosmetics.csv)의 일부 데이터세트에 기반한다. 이 상점에서는 매장 상품진열과 교차판매 판촉 직원의 훈련, 그리고 교차판매를 증진시키는 궁극의 구매시점 전자추천 시스템을 구축하기 위하여 이러한 아이템들의 구매 간 연관성을 분석하고자 한다. 우선 <표 14.14>에 나타난 이진행렬 형식의 일부 데이터만을 고려하자.**

**a. 행렬의 여러 값들을 선택하여 그 의미를 설명하시오.**



* 4번째 행의 의미는 고객이 ‘Nail Polish, Brushes, Concealer, Bronzer’ 를 같이 구매한다는 것을 의미합니다. 또한 9번째 행의 의미는 ‘Concealer’만 구매하는 것을 의미하며, 11번째 행은 ‘Nail Polish’를 구매하면 ‘Bronzer’도 함께 구매한다는 것을 의미합니다.

**b. <표 14.15>에 나타난 연관규칙 분석 결과를 보고, 다음에 답하시오.**



1. **첫 번째 행에서 “신뢰도” 출력의 의미와 어떻게 계산되는지 설명하시오.**

* 신뢰도는 confidence를 의미합니다. 즉 ‘Blush, Concealer, Mascara, Eye shadow, Lipstick’이 함께 있는 거래가 A개일때, 이 A개 거래 중에 ‘Eyebrow Pencils’가 포함된 거래가 B개라면 “IF Blush, Concealer, Mascara, Eye shadow, Lipstick이 함께 판매되면 THEN Eyebrow Pencils도 같은 거래에서 판매된다”는 연관규칙이 성립합니다. 따라서 신뢰도는 B / A로 계산할 수 있습니다.

1. **첫 번째 행에서 “지지도” 출력의 의미와 어떻게 계산되는지 설명하시오.**

* 지지도는 support를 의미합니다. 즉 거대 체인 화장품점의 총 N개의 거래 중, 조건부 항목집합(Blush, Concealer, Mascara, Eye shadow, Lipstick)과 결론부 항목집합(Eyebrow Pencils)을 동시에 포함하는 B개의 거래 확률을 나타냅니다. 따라서 지지도는 B / N로 계산할 수 있습니다.

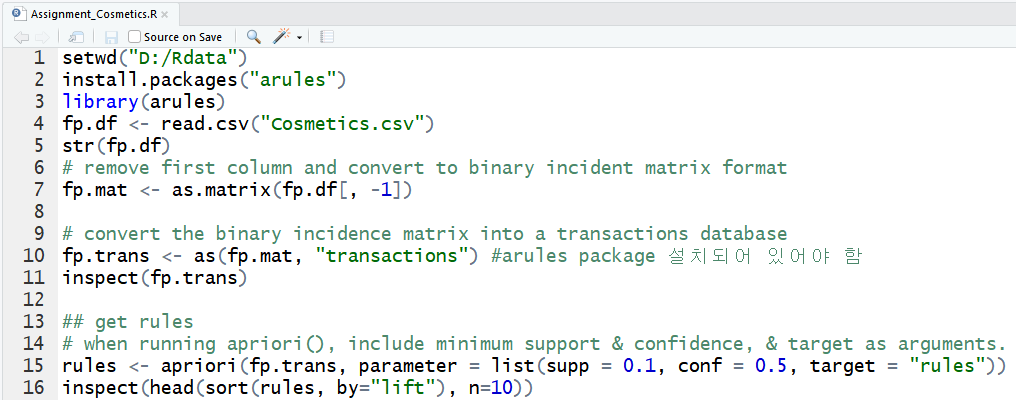
1. **첫 번째 행에서 “향상”의 의미와 어떻게 계산되는지 설명하시오.**

* 향상은 lift를 의미합니다. 연관성의 강도를 측정하기 위한 측정도구인 향상도는, 신뢰도를 기준 신뢰도로 나눈 값을 나타냅니다. 이 때 기준신뢰도는 거래 내의 결론부 항목집합이 각 규칙의 조건부와 독립적이라고 가정하여 얻어진 것이므로, 결론부가 발생할 확률과 동일하게 됩니다. 따라서 향상도는 신뢰도(0.302) / 결론부(Eyebrow Pencils)가 판매될 확률로 계산할 수 있습니다.

1. **첫 번째 행에서 그 규칙이 표현하는 의미를 말로 설명하시오.**

* 첫 번째 행이 의미하는 바는, ‘Blush, Concealer, Mascara, Eye shadow, Lipstick’을 함께 구매한 고객은 ‘Eyebrow Pencils’도 구매한다는 것을 나타냅니다.

**c. 이제 (Cosmetics.csv 파일에 저장된) 화장품 구매에 대한 전체 데이터세트를 사용하시오. R을 사용하여 이 데이터에 연관규칙을 적용하시오. (기본 파라미터를 사용할 것) [지지도 0.1, 신뢰도 0.5 가정]**

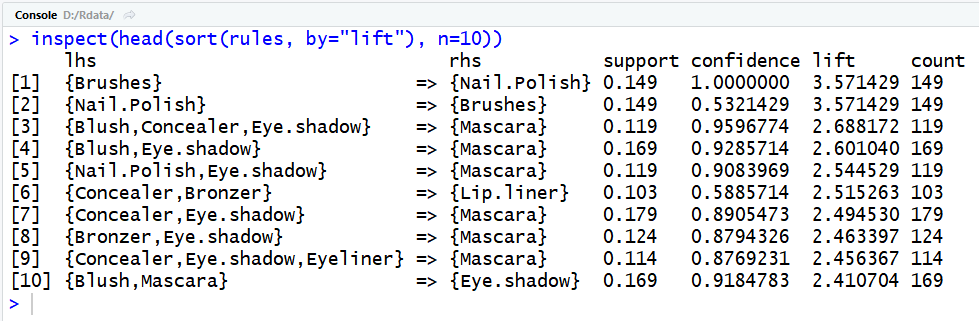


* Apriori 알고리즘을 실행하는 R코드를 참고하였습니다. (교재 p.364)

1. **출력된 처음 세 개의 규칙들을 말로 해석하시오.**

**(support가 높으면 rule이 단순해 진다.)**

**(count는 lhs와 rhs가 동시에 들어가 있는 수)**



* 공통) count에 따라 support와 confidence가 변경되며, 위 결과만 볼 때 lift가 모두 1보다 크므로 현재규칙이 결론부 항목집합을 찾는데 유용한 규칙임을 알 수 있습니다. 즉 향상도가 클수록 연관성은 더 강해집니다.
* 첫 번째 행은 ‘Brushes’를 구입하면 ‘Nail Polish’도 구입한다. 라는 연관성을 알 수 있습니다.
* 두 번째 행은 ‘Nail Polish’를 구입하면 ‘Brushes’도 구입한다. 라는 연관성을 알 수 있습니다.
* 세 번째 행은 ‘Blush, Concealer, Eye shadow’를 구입하면 ‘Mascara’도 구입한다. 라는 연관성을 알 수 있습니다.