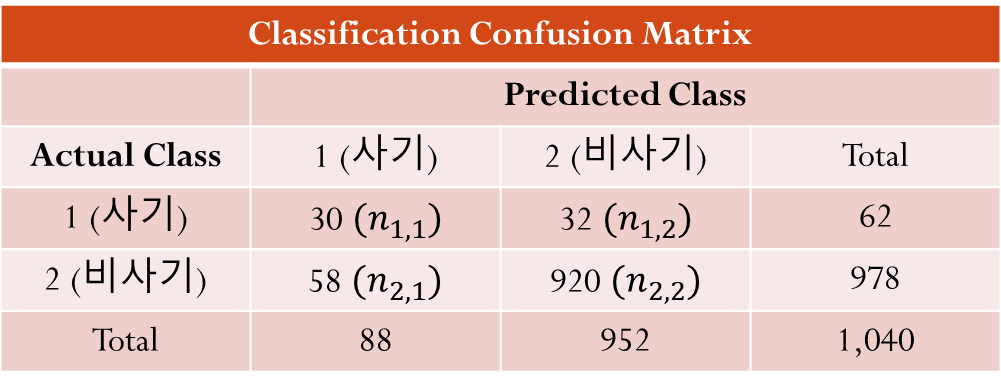
**[세 번째 과제]**

**IT응용시스템공학과**

**1494053**

**김희택**

**5.1 거대 데이터세트에 데이터마이닝 과정이 적용되어 88개의 레코드들을 사기라고 분류하고(그 중 30개가 올바름) 952개를 비사기라고 분류하였다. (그 중 920개가 올바름). 정오행렬을 작성하고 전체적인 오차율을 계산하시오.**



* 정오행렬은 위와 같다.
* 오차율(Error Rate)은 (잘못 분류된 레코드의 총합) / (전체 레코드) 이므로 (58 + 32) / (1040) = 약 0.0865 이다. 따라서 약 8.65%가 오차율이다.

**5.2 이 과정에 사기라고 분류된 레코드들의 비율을 변경할 수 있는 조정 가능한 컷오프(임계값) 메커니즘이 있다고 하자. 컷오프 값을 올리거나 내리면 어떤 효과가 있는지 다음에 대해 설명하시오.**

**a. 정말 사기인 레코드에 대한 분류 오차율 (컷오프 값 0.5 가정)**

* 1 - Sensitivity를 말한다.
* 1 - =
* 컷오프 (임계값)을 올리게 될 경우, 사기가 비사기로 분류되는 경우 가 많아지게 되고 동시에 사기가 사기로 분류되는 경우 는 줄어든다.
* 따라서 ‘정말 사기인 레코드에 대한 분류 오차율’ 은 증가한다.
* 컷오프 (임계값)을 내리게 될 경우, 사기가 사기로 분류되는 경우 가 많아지게 되고 동시에 사기가 비사기로 분류되는 경우 는 줄어든다.
* 따라서 ‘정말 사기인 레코드에 대한 분류 오차율’ 은 감소한다.

**b. 정말 비사기인 레코드에 대한 분류 오차율 (컷오프 값 0.5 가정)**

* 1- Specificity를 말한다.
* 1 - =
* 컷오프 (임계값)을 올리게 될 경우, 비사기가 사기로 분류되는 경우 가 줄어들게 되고 동시에 비사기가 비사기로 분류되는 경우 는 많아진다.
* 따라서 ‘정말 비사기인 레코드에 대한 분류 오차율’ 은 감소한다.
* 컷오프 (임계값)을 내리게 될 경우, 비사기가 비사기로 분류되는 경우 가 줄어들게 되고 동시에 비사기가 사기로 분류되는 경우 는 많아진다.
* 따라서 ‘정말 비사기인 레코드에 대한 분류 오차율’ 은 증가한다.

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

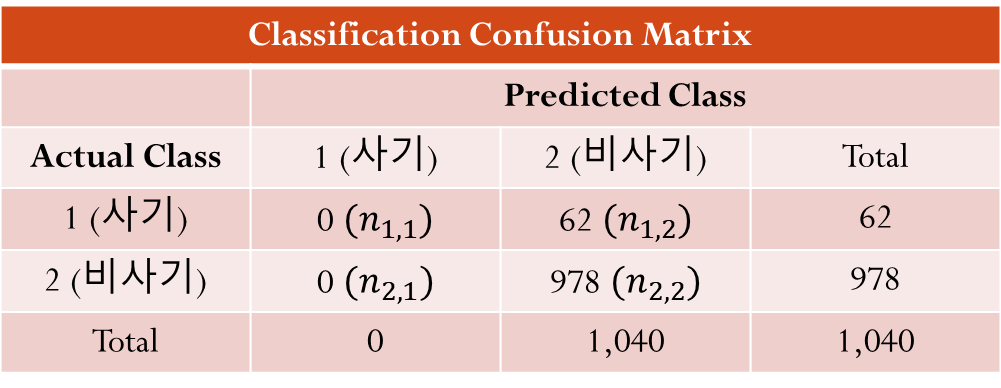
**a. 왼쪽에서 첫 번째와 두 번째 막대의 의미를 해석하시오.**

* 왼쪽에서 첫 번째 막대는 ‘가장 가능성 있는 클래스1’으로 모형이 순위 매긴 레코드들의 10%를 취함으로, 10% 레코드들의 랜덤선택 중의 1의 개수 보다 클래스1에 속하는 레코드를 약 6.5배 더 얻을 수 있다.
* 왼쪽에서 두 번째 막대는 위에서 취한 상위 10%를 제외하고 그 다음 10%를 취함으로, 10% 레코드들의 랜덤선택 중의 1의 개수 보다 클래스1에 속하는 레코드를 약 3배 더 얻을 수 있다.

**b. 이 정보를 실제로 어떻게 사용할 수 있는지 설명하시오.**

* X축 각 10분위당 주어지는 레코드 수는 1040 / 10 = 104 이다. 10%마다 클래스1의 비율을 비교.
* 대부분의 관심 있는 목표집단을 왼쪽에서 첫 번째와 두 번째 막대를 선택하므로 판단할 수 있다.

**c. 또 다른 분석가는 모든 것을 비사기라고 분류함으로써 모델의 정확도를 개선할 수 있다고 주장한다. 그렇게 한다면, 오차율은 무엇인가? (5.1의 정오분류표 이용)**



* 정오행렬은 위와 같다.
* 오차율(Error Rate)은 (잘못 분류된 레코드의 총합) / (전체 레코드) 이므로 (0 + 62) / (1040) = 약 0.0596 이다. 따라서 약 5.96%가 오차율이다.
* 초기 (약 8.65%) 보다 약 2.69% 오차율이 감소하였다.

**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.