5.1 거래 데이터 세트에 데이터마이닝 과정이 적용되어 88개의 레코드를 사기라고 분류하고 (그중 30개가 올바름) 952개를 비사라기라고 분류하였다(그중 920개가 올바름). 정오행렬을 작성하고 전체적인 오차율을 계산하시오.

	Predicted Class	
Actual Class	사기(C_1)	비사기(C_2)
사기(C_1)	30($n_{1,1}$)	32($n_{1,2}$)
비사기(C_2)	58($n_{2,1}$)	920($n_{2,2}$)

$err = (58+32)/1040 = 8.65\%$

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.

▶ 컷오프 값이 증가하면 C_1으로 분류되는 값이 감소한다.

a. 정말 사기인 레코드에 대한 분류 오차율(1-sensitivity)

컷오프를 올리면 정말 사기인 레코드에 대한 분류 오차율(1-sensitivity)이 증가한다.
(컷오프를 올리면 $n_{1,1}$의 값은 감소하고 $n_{1,2}$의 값은 증가하기 때문에

$1-sensitivity = \frac{n_{1,2}}{n_{1,1}+n_{1,2}}$의 값이 증가한다. 분모의 값은 고정이다.)

b. 정말 비사기인 레코드에 대한 분류 오차율(1-specificity)

컷오프를 올리면 정말 비사기인 레코드에 대한 분류 오차율(1-specificity)이 감소한다.
(컷오프를 올리면 $n_{2,1}$의 값은 감소하고 $n_{2,2}$의 값은 증가하기 때문에

$1-specificity = \frac{n_{2,1}}{n_{2,1}+n_{2,2}}$의 값이 감소한다. 분모의 값은 고정이다)

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

a. 왼쪽에서 첫 번째와 두 번째 막대의 의미를 해석하시오.

왼쪽에서 첫 번째 막대는 관심 있는 Class(사기 거래)에 속할 확률 높은 상위 10%의 부분집합이다. 임의로 레코드를 10% 선택하는 것에 비해 관심 있는 Class에 속할 가능성이 약 6.5배 높다.

왼쪽에서 두 번째 막대는 관심 있는 Class(사기 거래)에 속할 확률이 높은 상위 10%~20%의 부분집합이다. 임의로 레코드를 10% 선택하는 것에 비해 관심 있는 Class에 속할 가능성이 약 3배 높다.

b. 이 정보를 실제로 어떻게 사용할 수 있는지 설명하시오.

세금 사기를 찾기 위해 소득 신고서를 조사할 때, 어떤 신고서를 얼마나 많이 조사해야 하는가를 결정할 때 사용한다.

c. 또 다른 분석가는 모든 것을 비사기라고 분류함으로써 모델의 정확도를 개선할 수 있다고 주장한다. 그렇게 한다면, 오차율은 무엇인가?

	Predicted Class	
Actual Class	사기(C_1)	비사기(C_2)
사기(C_1)	0($n_{1,1}$)	62($n_{1,2}$)
비사기(C_2)	0($n_{2,1}$)	978($n_{2,2}$)

$err = 62/1040 = 5.96\%$

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

모든 것을 비사기로 분류할 때 오차율이 감소 했지만 향상도는 보장하지 못한다. 모형에서 둘의 관계를 파악하고 필요에 따라 적절한 오차율과 향상도를 가지는 기준값 설정이 필요하다.