4.4 이익도표

이익도표(lift chart)는 주어진 분류기의 예측능력을 표 및 그림으로 나타내는 방법이다. 이익도표를 그리는 절차는 다음과 같다.

1. 모형적합을 통하여 사후확률을 계산한다.
2. 사후확률이 큰 순서에 따라 자료를 정렬시킨다.
3. 정렬된 자료를 균일하게 K등분한다.
4. K등분의 각 등급에서 출력변수의 $y = 1$의 클래스에 대한 빈도를 구한다.
5. K등분의 각 등급에서 %Captured Response, %Response 및 Lift 통계량을 구한다.
6. 수평축에는 K등분의 등급을, 수직축에는 위의 3개의 통계량 중 하나를 이용해 그래프를 그린다.

이익도표에 사용되는 통계량들은 다음과 같이 정의된다.

- $\%\text{Captured Response} = \frac{\text{해당 등급에서 출력변수가 y=1인 빈도}}{\text{전체 자료에서 y=1의 빈도}} * 100\%$
- $\%\text{Response} = \frac{\text{해당 등급에서 출력변수가 y=1인 빈도}}{\text{해당 등급의 자료의 수}} * 100\%$
- $\text{Baseline Lift} = \frac{\text{전체 자료에서 출력변수가 y=1 클래스의 자료수}}{\text{전체 자료수}} * 100\%$
- $\%\text{Captured Response} = \frac{\text{해당 등급에서 출력변수가 y=1인 빈도}}{\text{전체 자료에서 y=1의 빈도}} * 100\%$

%Captured response는 해당 등급에서의 민감도, %Response는 해당 등급에서의 정분류율이며, Lift는 모형을 사용하지 않고 사전확률로만 분류할 때에 비하여 모형을 사용하였을 때의 정분류율의 증가비를 나타낸다. 표 4.2는 이익도표의 예제이다. 전체 자료수는 2000개이고 이 중 y=1인 클래스의 자료수는 381개이다. 전체자료를 사후확률을 기준으로 10개의 등급을 만들고 각 등급에서 관련 통계량을 계산하였다. 첫번째 등급의 Lift는 4.57로 첫번째 등급에 속하는 자료는 모형을 사용하지 않았을 때에 비하여 모형을 사용하였을 때 4.57배의 정분류율의 증가가 있었다는 것을 알 수 있다. 실제 자료분석에서는 Lift가 가장 많이 쓰이는 통계량이며 $P(y = 1|x)$가 상대적으로 높은 등급을 선택할 때 Lift값을 기준으로 나눈다. 예를 들면 Lift가 2 이상이면 $P(y = 1|x)$가 상대적으로 높은 등급이라고 한다면 2등급까지의 자료가 대상이 된다. 목표마케팅에서는 이러한 그룹만을 대상으로 마케팅 작업을 수행한다. 그림 4.4는 표 4.2에 있는 Lift 값을 도표로 그린 것이다.

그림 4.4: 이익도표

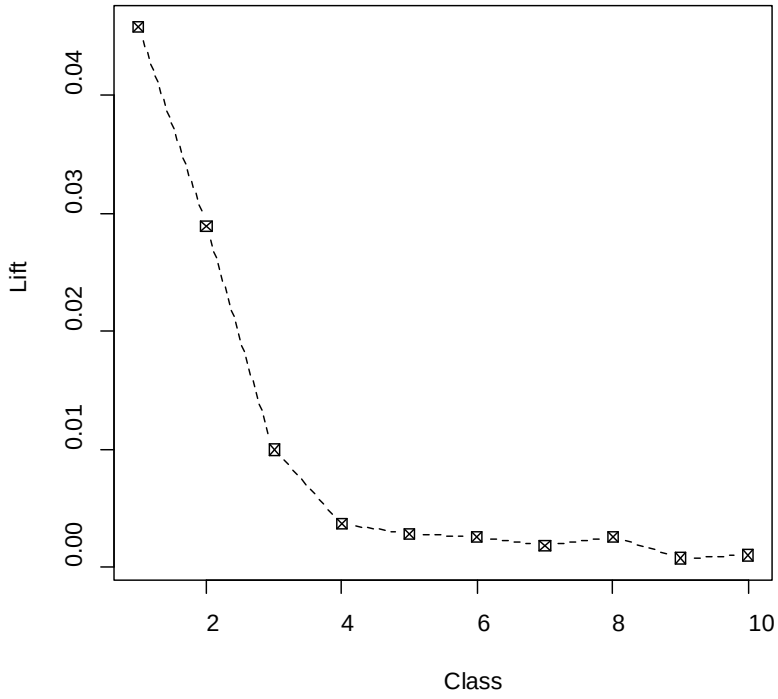

표 4.2: 이익도 테이블

등급	범주1의 빈도	%Captured response	%Response	Lift
1	174	174/381=45.6	174/200=87.0	87.0/19=4.57
2	110	110/381=28.8	110/200=55.0	55.0/19=2.89
3	38	38/381=9.9	38/200=19.0	19.0/19=1.00
4	14	14/381=3.6	14/200=7.0	7.0/19=0.36
5	11	11/381=2.8	11/200=5.5	5.5/19=0.28
6	10	10/381=2.6	10/200=5.0	5.0/19=0.26
7	7	7/381=1.8	7/200=3.5	3.5/19=0.18
8	10	10/381=2.6	10/200=5.0	5.0/19=0.26
9	3	3/381=0.7	3/200=1.5	1.5/19=0.07
10	4	4/381=1.0	4/200=2.0	2.0/19=0.1
Base Line Lift = 381/2000 = 19.05				