데이터마이닝(04/09)

# 5장 과제 (김광래, 박동민, 홍진영)

**5.1.** 거래 데이터세트에 데이터마이닝 과정이 적용되어 88새의 레코드들을 사기라고 분류하고(그중 30개가 올바름) 952개를 비사기라고 분류하였다.(그중 920개가 올바름). 정오행렬을 작성하고 전체적인 오차율을 계산하시오.

우선, 88개의 레코드를 ‘사기’라고 분류했다는 말은 88개의 레코드를 ‘사기’로 ‘예측’했다는 뜻이다. 그리고 괄호 내의 ‘그 중 30개가 올바르다는 뜻은 ‘사기’로 ‘예측’된 레코드 중, ‘실제’로 ‘사기’인 레코드는 30개라는 뜻이다. 그렇다면, 사기(예측)-사기(실제) = 30/88이고,   
사기(예측) – 비사기(실제) = 58/88 이라는 소리다.  
  
반대로, 952개를 비사기로 분류했다는 말은 ‘예측’한 레코드 중 952개를 ‘비사기’로 골랐다는 뜻이고, ‘그 중 920개가 올바르다’는 뜻은 ‘예측’된 ‘비사기’ 중 ‘실제’ ‘비사기’의 개수가 920개 라는 뜻이다.  
그렇다면, 비사기(예측) – 비사기(실제) = 920/952, 비사기(예측) – 사기(실제) = 32/952 이라는 소리다.

[개수로 구분한 정오행렬]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **실제** | | 총계 |
| **예측** |  | 사기 | 비사기 |  |
| 사기 | 30 | 58 | 88 |
| 비사기 | 32 | 920 | 952 |
| 총계 |  | 62 | 978 | 1040 |

[퍼센트로 구분한 정오행렬]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **실제** | | 총계 |
| **예측** |  | 사기 | 비사기 |  |
| 사기 | **0.341**(30/88) | **0.659**(58/88) | 1 |
| 비사기 | **0.034**(32/952) | **0.966**(920/952) | 1 |
| 총계 |  |  |  |  |

정오행렬로 정리한 결과, 사기와 비사기에 대한 오차율을 파악할 수 있었다.

‘사기’에 대한 오차는 58건, ‘비사기’에 대한 오차는 32건. 고로, 총 1040건의 레코드 중, 90건의 오차가 발생하여 90/1040 = **0.0865** 의 오차율이 존재한다.

**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 대해서 설명하시오.

|  |  |
| --- | --- |
| 예측(비사기), 실제(사기) **[0.034]** | 예측(사기), 실제(사기) **[0.341]** |
| 예측(비사기), 실제(비사기) **[0.966]** | 예측(사기), 실제(비사기) **[0.659]** |

현 상황에 대해 그림으로 묘사하면 다음과 같다.

**Cutoff**

그리고 현 상황에서 Cutoff를 올리면 다음과 같다.

**Cutoff**

|  |  |
| --- | --- |
| 예측(비사기), 실제(사기) **[?]** | 예측(사기), 실제(사기) **[?]** |
| 예측(비사기), 실제(비사기) **[?]** | 예측(사기), 실제(비사기) **[?]** |

|  |  |
| --- | --- |
| 예측(비사기), 실제(사기) **[?]** | 예측(사기), 실제(사기) **[?]** |
| 예측(비사기), 실제(비사기) **[?]** | 예측(사기), 실제(비사기) **[?]** |

Cutoff를 내리면 다음과 같다.

**Cutoff**

1. 정말 사기인 레코드에 대한 분류 오차
2. Cutoff를 올릴 때

우선 생각해보면 ‘사기’를 기준으로 Cutoff를 움직일 때, 사기(예측) –사기(실제)의 분류 개수는 줄어들 것이다. 고로, ‘사기’에 대한 민감도가 떨어질 것이라는 소리다.  
  
반면, 사기(예측) – 비사기(실제)인 분류, 즉 ‘민감도에 대한 오차율’은 증가할 것이다. 고로, Cutoff를 높일 경우에 **정말 사기인 레코드에 대한 분류 오차율은 증가할 것이다.**

1. Cutoff를 낮출 때

‘사기’를 기준으로 Cutoff를 낮출 때, 사기(예측)-사기(실제)의 분류 개수는 증가할 것이다. 예측을 통해 분류된 데이터 중, 실제 사기로 분류된 데이터가 증가한다는 소리다. 고로 ‘사기’에 대한 민감도는 증가한다.

반면, 사기(예측) – 비사기(실제)인 분류, 즉 ‘민감도에 대한 오차율’은 감소할 것이다. 고로, Cutoff를 낮출 경우에 **정말 사기인 레코드에 대한 분류 오차율은 감소할 것이다.**

**b)** 정말 비사기인 레코드에 대한 분류 오차율

1. Cutoff를 올릴 때

‘사기’에 대한 Cutoff를 올릴 때, 예측을 ‘비사기’로 하는 영역의 범위는 증가한다. 범위가 증가하면 할수록 ‘비사기’로 예측된 공간 내의 실제 ‘비사기’ 데이터 개수도 증가한다는 뜻이다. 고로, **Cutoff를 올릴 때, ‘비사기’의 예측, 특이도는 증가한다.**

반면, ‘비사기’의 예측 영역이 증가한다는 것은 실제 ’사기’로 분류된 레코드도 ‘비사기’로 예측된 영역에 포함될 가능성이 있다는 것이다. 하여, **특이도의 오차율은 증가한다**.

1. Cutoff를 내릴 때

‘사기’에 대한 Cutoff를 내릴 때, 즉 사기로 예측할 영역을 늘릴 때, 예측이 ‘비사기’인 영역은 줄어든다. 고로, 예측이 ‘비사기’인 영역 자체가 줄기 때문에 예측-실제가 함께 ‘비사기’인 개수도 줄어든다. **Cutoff를 내릴 때, 비사기에 대한 특이도는 감소한다.**

반면, ‘비사기’로 예측할 영역이 줄어들기 때문에 실제 ‘비사기’인 데이터도 ‘사기’로 예측된 영역에 포섭된다. 그래서 **Cutoff를 내릴 때, 비사기에 대한 특이도의 오차율은 증가한다.**

**5.3**

****

****

**출처:** <https://www.congress.gov/>

미 115대 연방의회 자료에서 확인하여 보면 전체 법안(Bills) 중 상/하원을 통과하고 대통령 서명에 의해 법안으로 확정된 안건(Became Law)은 11,206건 중 417건, 전체의 약3.7% 다. 즉, 우리의 **‘나이브 벤치마크’는 3.7%이다.**

**가능한 오분류의 유형**으로는 ‘안건이 법안이 되지 못했는데 법안이 될 것이라고 예측한’ 경우와 ‘법안이 될 수 있는데 법안이 되지 못할 것이라’ 예측하는 2가지 유형이 있다.

측도로 전체적인 정확도를 활용하기엔 너무 낮다고 할 수 있는 수치인 3.7%이며, 다른 가능한 측도로는 법안 발의 정당과 집권당의 일치여부, 법안의 주제 등을 활용할 수 있을 것으로 보인다.

해당 측도를 통해 어떤 변수를 넣으면 '3.7%’보다 높은 통과율을 보일 수 있는지 찾아보려 한다. 생각해볼 수 있는 측도 중 대표적으로 법안 발의 정당과 집권당의 일치여부, 법안의 주제 등을 활용할 수 있을 것으로 보인다.



우리는 연방의회의 법안 통과율에 ‘집권당의 유무’가 변수가 될 수 있다고 보았다. 그것을 입증하기 위해 법안을 발의한 의원 중 그들의 정당을 변수로 데이터를 보았다. 해당 그림은 115대 연방의회에서 발의된 법안 중 Republican이 5,645명, Democratic이 5,509명임을 말하는 자료다. 덧붙여, Independent는 52명임을 말하고자 한다.



위 그림은 발의된 법안 중 'Became Law’, 즉 통과된 법안을 보았을 때 발의자의 정보를 본 것이다. Republican은 297명, Democratic은 120명이다. 115대 연방의회 당시 집권당이었던 공화당은 약 **5.2%**의 발의안이 승인되었고, 민주당은 약**2.2%**의 발의안이 승인되었다.

나이브 벤치마크와 함께 얘기하자면, 공화당은 5.2%의 통과율을 보여 나이브 벤치마크를 상회하고, 민주당은 나이브 벤치마크 아래 쪽으로 분포한다. 하여, ‘향상차트를 만들기 위한 조건으로 ‘집권당’ 변수는 좋은 측도가 된다.



법안 통과율을 위한 두 번째 변수로 법안의 ‘주제’를 살펴보았다. 위 그림 중 상단에 위치한 그림은 주제를 중심으로 보았을 때 ‘발의’된 법안, 하단에 위치한 그림은 ‘통과’된 법안이다.

주제별로 통과율을 살펴보는 경우 Government Operations and Politics의 경우, 119/890 = 약 13.37%, International Affairs의 경우 28/481= 약 5.82%로 ‘나이브 벤치마크’인 3.7%보다 높은 통과율을 보인다.

반면, ‘Health’의 경우, 26/1,307 = 1.99%의 통과율을 보였다. 해당 통과율은 ‘나이브 벤치마크’보다 낮은 통과율을 보이기에 향상차트를 적용할 때 제외될 수 있는 항목이다.

위와 같은 방식으로 발의된 법안 대비 통과된 법안을 파악하려면 ‘나이브 벤치마크’를 통해 기준점을 세우고, 해당 기준 대비 상향 확률을 보이는 변수들을 파악하면 통과율을 예측하기 수월해진다.

이러한 점으로 보아 통과율을 예측할 때, 단일 측도만을 사용하기보단 각각의 단일변수들을 모아 여러 측도를 중복으로 사용하면 법안 통과율을 보다 좀 더 정밀하게 예측해볼 수 있다고 생각한다.

법안에 영향을 받는 사업을 진행한다고 하는 경우, 예측을 통해 안건이 통과될 확률을 경향이라고 하면 법안이 될 경향이 가장 큰 안건부터 사업 계획에 적용하여 나갈 수 있게 된다. 즉 경향의 잠재적인 역할이란 제한된 자원 하에서 의사결정을 가장 효율적으로 내릴 수 있게 해주는 지표의 역할을 한다.

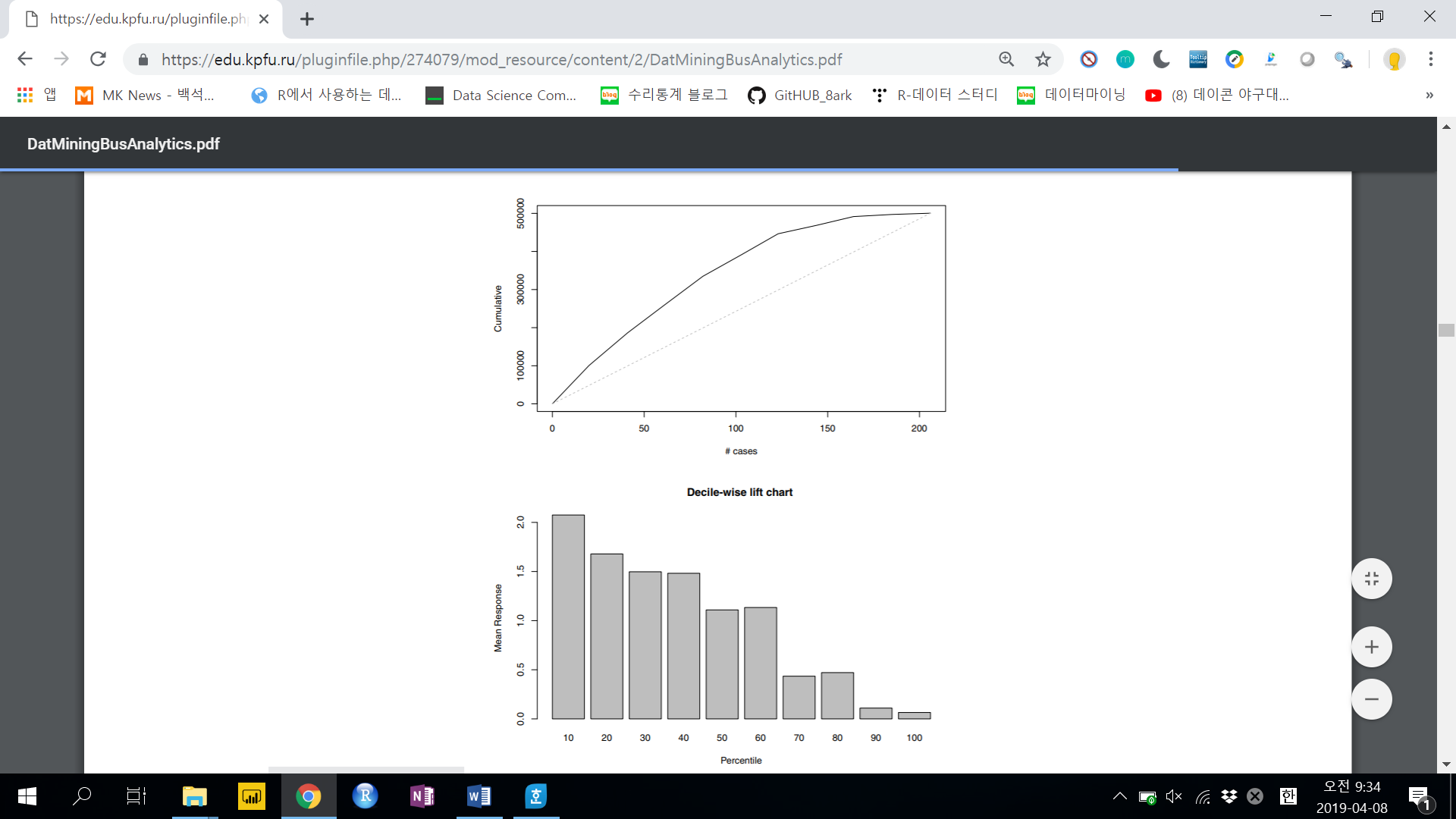
**5.6.**

1. 문제가 말하고자 하는 것은 ‘나이브 벤치마크’로 삼을 값을 설정하는 것이다. 다른 변수들을 생각하지 않고 1인당 평균이익을 기준으로 1,000명에 대한 추정 수익을 잡으면 그 값은 1,000 x 2,128$ = **2,128,000$** 이다. 즉, 2,128,000$는 나이브 벤치마크를 통해 얻은 소프트웨어 서비스의 판매 추정액이다.
2. 판매 영업 비용의 두 배 이상을 평균 수익으로 내고 싶다는 말은 ‘영업 이익’에 대한 1인당 평균 이익을 2,500$ 로 만들고 싶다는 이야기다. 영업 비용은 기본적으로 2,500$씩 들고, 그 두 배의 수익은 5,000$ 이다. 고로 영업 비용은 제외한 두 배 이상의 수익인 2,500$ 이다.

고로 소프트웨어 회사는 향상차트를 통해 판매가능한 가격을 내림차순으로 정렬한 뒤, 상위 10%부터 차근차근 ‘나이브 벤치마크’와 비교한 평균응답률을 보면 된다. 우선적으로 나이브 벤치마크와 비교했을 때 1인당 평균 이익이 2배가 되는 지점은

2,500/2,128 = 1.175으로 1.175가 되는 응답지점을 찾으면 된다.

165 페이지의 그림 상에는 Y축에 해당하는 평균 응답, 즉 ‘나이브 벤치마크’와 비교한 향상차트의 판매가격의 비율이 존재한다. 해당 그래프에서 1.175보다 큰 지점을 살펴보면 된다. 그림 상에는 1.175보다 크다고 느껴지는 지점이 누적 상위 60%까지이다. 고로, 소프트웨어 회사는 상위 누적 60%까지만 판매를 진행하면 ‘나이브 벤치마크’와 비교했을 때 2배 이상의 수익률을 올릴 수 있다.



1.175

**HW\_3\_DataMining(5장)**

**수업: 데이터마이닝**

**담당: 이성임 교수**

**학번 및 이름: 32121391 박동민**

**32150337 김광래**

**홍진영**