**비즈니스 모델링 2**

**Assignment 3**

**Due: 12/05, 2020 (토요일), 11:59PM**

**(Total Points: 22)**

**Instructions:** 다음 문제들을 잘 읽고, 요구되는 코드를 작성하고 설명하시오. 코드와 설명을 포함한 모든 답안을 ‘HW3\_김일주(자기이름).r’ 하나의 파일에 저장하여 제출하시오. 설명을 요하는 문제의 경우 주석을 사용하도록 하고, 코드와 설명이 몇번 문제에 대한 답인지를 반드시 표시하시오. 늦은 제출은 받지 않으니, 제출시한에 맞추도록 유의하시오. 반드시 자신들의 이해를 충분히 보이는 방법으로, 자신들의 Words를 사용하여 답안을 작성하시오. 질문있을 시 언제든지 [ijkim@mju.ac.kr](mailto:ijkim@mju.ac.kr)로 연락하시오.

1. **(3pts) A data mining routine has been applied to a transaction dataset and has classified 88 records as fraudulent (30 correctly so) and 952 as non-fraudulent (920 correctly so) (거래 데이터세트에 데이터마이닝 과정이 적용되어 88개의 레코드들을 사기라고 분류하고 (그중 30개가 올바른 분류) 952개를 비사기라고 분류하였다 (그중 920개가 올바름)). Construct the confusion matrix and calculate the overall error rate. (문제에서 제공한 정보를 바탕으로 confusion matrix를 작성하고 전체적인 오차율을 계산하시오. Confusion matrix에서 어떤 클래스가 무엇을 나타내는지 확실하게 명시하시오.)**

|  |  |  |
| --- | --- | --- |
| **Actual class**  **Predict class** | **사기 맞음(0)** | **사기가 아님(비사기)(1)** |
| **사기 예측(0)** | **30** | **58** |
| **비사기 예측(1)** | **32** | **920** |

**여기서 예측이라고 작성된 부분(row)이 predict class이며, 나머지 부분(column)이 actual class이다. 오차율은 (58+32)/1040 이므로 소수점 둘째자리까지 나타내면 0.087%이다.**

1. **(3pts) Suppose that this routine has an adjustable cutoff(threshold) mechanism by which you can alter the proportion of records classified as fraudulent. Describe how moving the cutoff up or down would affect (이 과정에 사기라고 분류된 레코드들의 비율을 변경할 수 있는 조정가능한 컷오프 (임계값) 메커니즘이 있다고 하자. 컷오프 값을 올리거나 내렸을 때 각각 다음의 경우 어떤 효과가 있을지에 대해 설명하시오. 작성한 답안이 이해하기 어려우면 감점이니 이해하기 쉽게 자신의 이해를 보이도록 작성하시오. 예를 들어, confusion matrix을 통해 각 경우의 error rate의 식과 그 값이 어떻게 변화하는지를 통해 설명하는 것도 좋은 방법임.)**
   1. **The classification error rate for records that are truly fraudulent (실제 사기인 레코드에 대한 분류 오차율)**

**기존의 실제 사기인 레코드에 대한 오차율은 32/(30+32) = 0.52%이다(소수점 둘째자리까지 표시) 이때 사기에 대한 cutoff를 내리게 되면 사기라고 예측되는 값이 늘어나게 된다.**

|  |  |  |
| --- | --- | --- |
|  | **사기맞음(0)** | **비사기맞음(1)** |
| **사기예측(0)** | **35** | **70** |
| **비사기 예측(1)** | **27** | **908** |

**이 결과값처럼 나왔다고 했을 때 사기 레코드에 대한 분류 오차율은 27/62 = 0.43%로 오차율이 기존대비 0.08% 줄어든다. 즉 진짜 사기인 경우를 더 잘 찾을 수 있는 효과를 얻을 수 있다.**

**반대로 사기의 cutoff 값을 올리면 사기라고 예측하는 개수가 줄어들게 된다.**

|  |  |  |
| --- | --- | --- |
|  | **사기맞음(0)** | **비사기맞음(1)** |
| **사기예측(0)** | **27** | **48** |
| **비사기 예측(1)** | **35** | **930** |

**위의 그림처럼 결과값이 나왔다고 했을 때 오차율은 35/62 = 0.56% 로 기존대비 0.05% 오차율이 증가한다. 즉 진짜 사기인 경우를 기존보다 잘 찾을 수 없는 효과를 얻을 수 있다.**

* 1. **The classification error rate for records that are truly nonfraudulent (실제 비사기인 레코드에 대한 분류 오차율)**

**실제 비사기인 레코드에 대한 분류 오차율은 58/978 = 0.059%이다**

**a의 방식대로 사기에 대한 cutoff값을 내리면 반대로 비사기 cutoff값은 올라가게 된다.**

|  |  |  |
| --- | --- | --- |
|  | **사기맞음(0)** | **비사기맞음(1)** |
| **사기예측(0)** | **32** | **78** |
| **비사기 예측(1)** | **30** | **900** |

**그러면 비사기라고 예측하는 값이 줄어들게 된다. 위에 그림처럼 결과값이 나왔다고 할 때 오차율은 78/978 = 0.079%로 오차율이 기존보다 증가한다. 즉, 진짜 비사기가 아닌 경우를 정확하게 찾을 수 없다는 효과를 얻을 수 있다.**

**반대로 비사기의 cutoff값을 내렸다고 할 때 비사기라고 예측하는 값이 올라가게 될 것이다.**

|  |  |  |
| --- | --- | --- |
|  | **사기맞음(0)** | **비사기맞음(1)** |
| **사기예측(0)** | **25** | **38** |
| **비사기 예측(1)** | **37** | **940** |

**위의 그림처럼 결과값이 나왔다고 할 때 오차율은 38/978 = 0.038%로 오차율이 기존보다 감소한다. 즉 진짜 비사기인 경우를 더 잘 찾을 수 있다는 효과를 얻을 수 있다.**