이학석사학위논문

데이터 마이닝 기법을 이용한 건강보험 사기 탐지

Fraud Detection in Health Insurance Using Data Mining Techniques

2017년 8월

서울대학교 대학원
통계학과
이민아

초 록

본 논문에서는 근래에 문제 의식이 제기되고 있는 건강보험 사기를 통계적 모형에 기반하여 예측하려 한다. 이를 위해 첫째, 보험사가 보유하고 있는 고객 정보, 보험 계약 정보, 보험금 지급 정보, 보험 설계사 정보를 기반으로 하여 보험 사기자를 적발하는 데에 중요한 변수를 선택하고 특성 생성(feature engineering)을 통해 예측력을 높일 수 있는 변수를 구성한다. 둘째, 구성된 변수에 의사결정나무(decision tree), 신경망 모형(neural network), 부스팅(boosting) 등의 데이터 마이닝 기법과 로지스틱 회귀분석(logistic regression)을 적용하여 보험 사기자를 예측하는 모형을 구축하고 셋째, 적용한 기법들 간의 예측력을 비교한다. 오분류율, ROC 그래프, 리프트 차트(Lift Chart) 을 이용하여 기법들 간의 예측력 및 성능을 비교하였고 모형 중 가장 효과적으로 보험 사기를 적발할 수 있는 모형을 찾고자 하였다. 최종적으로 모형 간의 비교를 통해 선택된 모형은 부스팅(boosting) 모형으로 전체 오분류율 12.13%의 성능을 보여준다. 4개의 모형에서 공통적으로 중요도 상위 변수로 뽑힌 5 개의 변수 중 4 개의 변수가 특성 추출을 통해 새로 구성된 변수였으며 이렇게 선택된 변수들을 통해 보험 사기 현황에 대해 이해할 수 있다.

데이터 마이닝 기법 중에 보험 사기를 가장 잘 예측하는 모형을 찾고, 새로운 변수 구성을 통해 예측력을 높이는 변수를 찾은 데에 본

논문의 의의를 찾을 수 있다.

주요어 : 데이터 마이닝, 보험사기, 지도학습

학번 : 2015-20306

목 차

I. 서론 **1**

1.1 연구 배경 1

1.2 연구 목적 2

II. 보험사기 적발을 위한 지도학습 데이터 마이닝 기법 **4**

2.1 로지스틱 회귀모형 4

2.2 의사결정나무 6

2.3 부스팅 9

2.4 신경망 모형 10

2.5 평가 방법 12

2.5.1 오분류표 12

2.5.2 ROC 그래프 13

2.5.3 리프트 차트 14

III. 분석 데이터 **16**

3.1 보험사기 패턴에 따른 변수의 생성 17

3.2 변수 선택 20

IV. 분석 결과 **24**

4.1 로지스틱 회귀모형 24

4.2 의사결정나무 27

4.3 부스팅 29

4.4 신경망 모형 31

V. 모형 평가 **33**
5.1 오분류표 33
5.2 ROC 그래프 34
5.3 리프트 차트 34

VI. 결론 **37**

참고 문헌 **39**

Abstract **41**

그 림 목 차

그림 1. 부스팅 과정 9
그림 2. 신경망모형 10
그림 3. 보험사의 관계형 데이터베이스 17
그림 4. 보험설계사별 사기 보험청구 관련 건수 19
그림 5. 모자이크 그림과 커널 밀도추정 그림 21
그림 6. 보험사기 적발 의사결정나무 27
그림 7. 의사결정나무에서 변수의 중요도 28
그림 8. 부스팅의 평균제곱오차 30
그림 9. 신경망 모형의 평균제곱오차 32
그림 10. 평가용 데이터에 대한 4 개 지도학습 모형의 ROC 그래프 35
그림 11. 평가용 데이터에 대한 4 개 지도학습 모형의 리프트 차트 35

표 목 차

표 1. 오분류표 13

표 2. 보험사기 패턴을 반영하여 생성된 입력변수 20

표 3. 변수 선택과정에서 선정된 입력변수 22

표 4. 로지스틱 회귀모형 추정결과 25

표 5. 로지스틱 회귀모형의 오분류표 25

표 6. 의사결정나무에서 변수의 중요도 28

표 7. 의사결정나무의 오분류표 29

표 8. 부스팅 모형에서 선택된 변수의 중요도 30

표 9. 부스팅 오분류표 31

표 10. 신경망 모형의 오분류표 32

표 11. 추정된 4 개 지도학습 모형의 평가용 데이터 오분류표 34

표 12. 평가용 데이터에 대한 4 개 지도학습 모형의 ROC Index 35

제 1 장

서론

1.1 연구 배경

「보험사기방지 특별법」(2016.9.30. 시행) 제2조에 따르면 “보험사기행위”란 보험사고의 발생, 원인 또는 내용에 관하여 보험자를 기망하여 보험금을 청구하는 행위를 말한다. 금융감독원의 보험사기 적발 통계에 따르면 2015년 보험사기 적발금액은 6,549억원(적발인원 83,431명)으로 전년대비 금액기준 9.2%(552억원) 증가하였다고 한다. 보험 사기가 이처럼 증가하는 원인으로 경기 침체로 인한 생계형 보험사기의 급증을 들 수 있겠으나, 더 근본적인 원인으로는 보험 사기에 대한 낮은 처벌 수위 때문에 죄의식 없이 보험 사기를 저지르는 사회적 인식을 들 수 있다. 실제로 지난 2012 년 기준 징역형 선고 비율은 보험사기범이 13.7%로 일반 사기범(46.6%)보다 훨씬 낮았다.

보험 사기 행위는 보험 회사에만 손해를 입히는 것 같아 보이지만 이는 보험사의 성실한 계약자들의 보험료 인상을 초래할 뿐만 아니라 보험 제도 자체의 존립 기반을 위협한다(송윤아, 2010). 보험연구원에 따르면 2014년 한 해 동안 총 4조 5천억원, 가구당 23만원 및 1인당 8.9만원의 보험금 누수가 발생한 것으로 추정된다. 최근 이러한 보험사기에 대한 문제 의식이 상승되면서 지난 2016년 9월 30일부터 보험사기자에 대한 처벌 수위를 높이는 법안인「보험사기방지 특별법」이 시행되었다. 보험 사기를 줄이기 위하여 강화된 처벌과 더불어 성능

높은 보험 사기 적발 시스템 구축 또한 요구된다.

1.2 연구 목적

송윤아와 정인영(2011)에 의하면 보험 사기 적발은 전통적으로 보험사기 조사자의 경험이나 직감에 기반하여 이루어져 왔다. 하지만 알려지지 않은 보험 사기 패턴이 존재하고 보험 사기 수법이 점점 지능화되고 있기 때문에 보험 사기 적발은 쉽지 않은 문제이다. 특히, 경험에만 의존하여 점점 다양화되고 지능화되는 보험 사기를 적발하는 것은 한계점이 있다. 이는 첫째, 경험적 지식에 기반한 결정은 제한된 변수에 의존하여 결정이 이루어지지만 결정에 영향을 미치는 다른 변수들이 분명히 존재하기 때문에 이를 의사결정과정에서 빼놓는 오류를 범하게 된다. 둘째, 경험적 지식에 기반하여 결정을 내릴 수 있는 전문 인력을 양성하는 것은 오랜 시간과 투자가 필요하다. 셋째, 점점 더 불어나는 데이터를 인간의 뇌로 상호 연관 시키는 것이 불가능하기 때문에 방대한 양의 데이터 속에서 보다 정확한 정보를 얻기 위해서는 과학적인 알고리즘이 요구된다(Bologa et al., 2010). 이러한 과학적 알고리즘으로 제안할 수 있는 것은 데이터 마이닝 기법으로, 방대한 양의 정보들에 데이터 마이닝 기법을 적용함으로써 데이터의 보고(寶庫) 속에 숨겨진 정보를 찾아내고 보험 사기자를 보다 정확하게 판별할 수 있을 것이다.

Hassan and Abraham (2013)은 건강보험 사기 예측은 데이터 마이닝 기법에 기반한 분류(classification) 모형이 가장 많이 사용되었고 그 성능 또한 높다고 하였다. 분류는 지도학습의 대표적인 분야로써 분류 기법 중에 많이 사용된 모형으로 로지스틱 회귀모형과 의사결정나무

모형을 꼽을 수 있다. 본 논문에서는 분류 기법 중 높은 예측력을 보이는 부스팅 모형과 신경망 모형을 추가적으로 적용해보도록 한다.

제 2 장

보험사기 적발을 위한 지도학습 데이터 마이닝 기법

2.1 로지스틱 회귀모형

로지스틱 회귀모형은 독립 변수의 선형 결합을 통해 어떤 사건이 일어날 확률을 예측하는데 사용하는 회귀 모형이다. 로지스틱 회귀분석은 독립 변수와 종속 변수 사이의 관계를 함수로 나타낸다는 점에서 일반 회귀 분석과 동일하지만 종속 변수가 명목 척도로 측정된 범주형 변수인 경우에 사용된다는 점에서 차이가 있다. 본 논문에서는 종속변수가 Fraud(보험금 부당 청구건)와 Non-Fraud(보험금 정당 청구건) 두 가지 값만을 갖는 이진(binary) 종속변수이다. 따라서, 데이터가 입력되었을 때 해당 데이터의 결과값을 특정 기준에 따라 나누어 두 가지 변수(Fraud/Non-Fraud)로 분류(classification)가 이루어진다. 또, 본 논문에서는 보험 사기에 미치는 여러 가지 요인들, $x_1, x_2, \ldots, x_p$가 입력 변수로 사용된다. 따라서 다중 선형 회귀모형으로 나타낼 수 있고 이를 수식으로 표현하면 다음과 같다.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon, \quad y = \begin{cases} 0, & \text{(보험금 부당 청구건)} \\ 1, & \text{(보험금 정당 청구건)} \end{cases}$$

하지만 $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$는 범위 $[0,1]$를 벗어날 수 있고 선형 회귀모형의 일반적인 가정인 오차항의 등분산성과 정규성을 충족시키지 않는 문제점이 있다. 이러한 문제점을 해결하기 위한 방안으로, 연속성을 만족하고 증가함수이며 $[0,1]$ 사이의 값을 가지는 연결 함수(link function), $g(x)$를 통하여 모형화하는 것이 로지스틱 회귀모형이다. 이를 나타내면 다음과 같다.

$$\Pr[y = 1|\mathbf{x}] = g(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

연결 함수의 형태에 따라 여러 모형을 가지고 있으며 계산 상의 편리성으로 인하여 연결함수의 형태가 $g(x) = \exp a(x)/(1+\exp a(x))$ 인 로지스틱 모형이 가장 많이 사용된다.

로지스틱 함수의 회귀계수를 추정하기 위해서는 일반적으로 최대가능도추정(maximum likelihood estimation) 방법을 사용한다. 즉, 가능도함수

$$L(\beta) = \prod_{i=1}^{n} \Pr[Y = y_i|\mathbf{x}_i]$$

를 최대화, 또는 로그 가능도

$$\ell(\beta) = \sum_{i=1}^{n} \left(y_i \mathbf{x}_i' \beta - \log(1 + \exp(\mathbf{x}_i' \beta) \right) \tag{2.1}$$

를 최대화하는 β를 구한다. 일반적으로 (2.1)의 로그 가능도 함수는 Iterative Reweighted Least Squares(IRLS) 알고리즘을 이용하여 구할 수 있으며 그 과정은 Hastie et al. (2009)를 참조할 수 있다.

보험 사기 적발 모형 생성 시 사용될 로지스틱 회귀모형의 특징은 다음과 같다.

(1) 오즈비나 회귀계수를 통해 설명변수와 종속변수 간의 관계를 설명하는 유용한 정보를 제공한다. 신경망과 같은 분석은 결과로부터 이러한 해석적 정보를 얻기 쉽지 않다.

(2) 가능한 많은 설명변수를 분석에 포함시킬수록 좋은 추정을 얻을 수 있다. 하지만 불필요하거나 관련성이 없는 입력변수를 포함시키는 것은 모형의 일반성을 떨어뜨리고 모형의 불안정성의 원인이 될 수 있다.

(3) 회귀모형은 각 설명변수들 간의 독립성을 가정하고 있다. 일부 변수들 간의 교호작용(interaction)을 모형에 포함시켜 분석하는 것도 가능하나 회귀분석을 통해 유용한 교호작용을 탐색하는 것은 쉽지 않다. 이에 비해 의사결정나무와 같은 분석방법은 교호작용의 효과 및 비선형성을 자동적으로 찾아내는 알고리즘이라고 할 수 있다.

2.2 의사결정나무

의사결정나무(decision tree)는 입력된 값에 대하여 의사 결정 규칙을 적용하여 출력 값을 예측하고 분류하는 모형이다. 의사결정나무는 성장, 가지치기, 타당성 평가, 해석 및 예측 단계로 형성된다. 나무의 성장 단계에서는 노드의 불순도(impurity)를 카이제곱 통계량과 같은 척도에 의해 측정한다. 일반적으로 많이 사용되는 측도는 다음과 같다.

- 카이제곱 통계량: $X^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i^2}$가 최대가 되는 분리를 선택한다.

- 지니 지수: $Gini = 1 - \sum_{i=1}^{k} \left(\frac{\#\text{ of } O_i}{n} \right)^2$가 최소가 되는 예측변수와 분리를 선택한다.

- 엔트로피 지수: $I(t) = -\sum_{i=1}^{k} p(j|t) \times \log_2 p(j|t)$가 최소가 되는 분리를 선택한다.

여기서 E_i와 O_i는 각각 i-번째 클래스의 기대도수(expected frequency)와 관찰도수(observed frequency)를 나타내며, n과 k는 각각 관찰값의 총 수와 클래스의 수이다. 한편, 위 식에서 $I(t)$는 노드의 엔트로피 즉, 노드 t의 불순도(impurity)를 나타내고, 엔트로피 지수를 최소로 만들어주는 예측변수와 그 때의 분리를 선택한다.

가지치기 단계에서는 많은 노드의 수로 형성된 나무에서 적절하지 않은 마디를 제거하여 최종적인 예측모형인 의사결정나무를 만드는 과정이다. 노드의 수가 많아지게 되면 결과적으로 발생되는 규칙이 많아지게 되고 이를 적당한 시점에서 가지치기를 함으로써 가장 적절한 수의 규칙을 생성하게 된다. 오분류율이나 오차를 크게 할 위험이 높은 가지, 부적절한 추론규칙을 가지고 있는 가지 등 불필요한 가지들을 다음의 4가지 방법을 주로 사용하여 가지치기를 실행한다.

- Chi-Square 검정의 유의확률 값 (p-value)

- Entropy 또는 Gini 지수의 순수도

- 분리를 위한 각 노드의 최저 데이터 수

- 의사결정나무의 깊이 수준 (나무의 크기)

타당성 평가는 이익도표(gain chart), 위험도표(risk chart) 또는 시험자료를 이용하여 평가가 이루어진다. 해석 및 예측 단계는 형성된

나무모형을 해석하고 이에 기반하여 예측모형을 구축한 후에 예측에 적용하는 단계이다.

의사결정나무는 종속변수가 범주형 변수인지, 수치형 변수인지에 따라 분류나무(classification trees)와 회귀나무(regression trees)로 구분된다. 본 논문에서는 종속변수가 Fraud(보험금 부당 청구건)와 Non-Fraud(보험금 정당 청구건) 두 가지 값만을 가지는 범주형 변수이기 때문에 분류나무(classification trees)를 사용한다. 보험 사기 적발 모형 생성 시 사용될 의사결정나무의 특징은 다음과 같다.

(1) 분류나 예측의 근거를 가시적으로 제공해 주기 때문에 모형의 이해가 쉽다. 따라서 새로운 자료의 모형에 적합함을 통해 다양한 실무 분야에서 대중적으로 활용이 가능하다.

(2) 비모수적 모형으로 선형성, 정규성, 등분산성 등의 가정이 필요 없기 때문에 데이터 선정이 용이하다.

(3) 연속형이나 명목형 데이터 값들을 기록된 그대로 처리할 수 있고 데이터를 구성하는 속성의 수가 불필요하게 많을 경우에도 모형 구축시 분류에 영향을 미치지 않는 속성들을 자동으로 제외시키기 때문에 데이터 전처리 단계에서 소요되는 시간과 노력을 단축할 수 있다.

(4) 연속형 변수를 비연속적인 값으로 취급하기 때문에 정보의 손실이 발생하고 연속형 데이터를 처리하는 능력이 신경망이나 다른 통계기법에 비해 떨어지며 결과적으로 예측력도 감소한다.

(5) 하위 노드로 갈수록 표본이 작아지는 특성으로 모형 구축 시에 사용되는 표본의 크기에 민감하다. 따라서 정확한 모형을 구축하

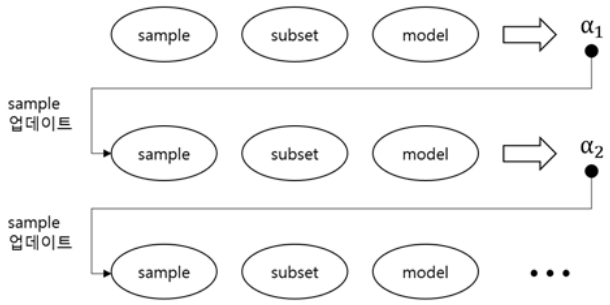


그림 1: 부스팅 과정

기 위해서는 서로 상이한 값을 갖는 데이터들을 충분히 확보해야 한다.

2.3 부스팅

부스팅(boosting)은 여러 개의 예측력이 약한 분류기(weak learner)들의 예측을 결합하여 예측력이 강한 분류기(strong learner)를 형성하여 분류의 정확도를 높이는 앙상블 기법 중에 하나이다. 부스팅 중에서도 본 논문에서 사용될 그래디언트 부스팅(gradient boosting)은 Freidman (2001) 에 의해 제안된 방법으로, 전체 데이터의 관측값에 같은 가중치를 준 상태에서 시작하여 정분류된 관측값에는 낮은 가중치를, 오분류된 관측값에는 높은 가중치를 주는 것을 반복하며 최종 분류기를 생성하는 방법이다. 그래디언트 부스팅의 알고리즘은 박창이 외 (2013)을 참조할 수 있다.

부스팅 모형의 특징은 다음과 같다.

(1) 신경망 모형과 마찬가지로 설명력이 약하다. 결과를 도출하는 과정을 명확하게 설명할 수 없다.

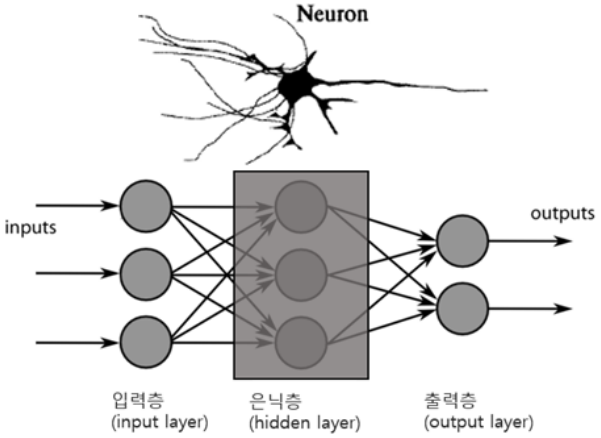


그림 2: 신경망모형

(2) 약한 분류기들을 결합함으로써 훈련 오차(training error)를 빨리 그리고 쉽게 줄일 수 있다.

(3) 2차, 3차 분류기에 들어가는 데이터는 기존 데이터의 일부만 적용되므로 training data 의 규모가 커야 한다.

2.4 신경망 모형

신경망 모형(neural network)는 인간의 뇌와 신경 시스템에서 착안하여 생물학적 프로세스 과정을 컴퓨터로 구현하기 위한 노력에서 시작되었다. 인간의 뇌와 같이 시냅스의 결합으로 네트워크를 형성한 인공 뉴런이 데이터에 기반한 반복적인 학습을 거쳐 입력값에 대한 최적의 출력값을 예측하는 통계학적 학습 알고리즘이다. 신경망 모형은 특히 방대한 데이터 속에 숨겨진 복잡한 패턴이나 연관 관계를 찾아내는 데에 유용하며 이를 기반으로 하여 예측이 필요한 다양한 응용분야에서 사용되고 있다. 그림 2은 다층신경망의 구조를 보여준다.

다층신경망은 입력층(input layer), 은닉층(hidden layer) 그리고 출력층(output layer)로 구성된다. 입력층은 각 입력변수에 대응되는 마디들로 구성되고 은닉층은 여러 개의 은닉 마디들로 구성되며 입력층으로부터 전달되는 변수값들의 선형 결합을 비선형 함수로 처리하여 출력층 또는 다른 은닉층에 전달한다. 출력층은 출력 변수에 대응하는 노드로, 명목형 목표 변수일 때는 클래스의 수만큼의 출력 노드를 생성한다. 보험 사기 적발 모형 생성 시 사용될 신경망 모형의 특징은 다음과 같다.

(1) 신경망 모형은 출력 데이터의 특성에 따라 연속형 변수일 때는 예측(prediction)을, 명목형 변수일 때는 분류(classification)를 실행한다.

(2) 훈련 과정을 거쳐서 변환이 필요하지만 연속형 자료와 범주형 자료 모두 처리할 수 있다.

(3) 신경망은 기존 통계적 기법과 비교하였을 때 높은 예측력을 보이는 강력한 분석기법이다. 비선형 모형에 결과값을 적용시킴으로써 최선의 값을 선택하므로 복잡한 구조의 데이터에서도 좋은 결과를 낼 수 있다.

(4) 설명력이 약하다. 결과를 도출하는 과정을 명확하게 설명할 수 없는 것이 신경망의 최대 단점이다. 따라서, 신경망은 결과의 도출과정보다 결과값 자체가 중요한 경우에 주로 사용되고 해석이 중요한 분야에서는 사용되지 않는다. 도출 과정은 알 수 없지만 판별력 분석으로 입력 변수의 중요도 순서는 알 수 있다.

(5) 많은 가중치를 추정해야 하므로 과대 적합의 문제가 자주 발생한

다. 과대적합을 피하기 위해서는 알고리즘을 조기 종료 시키는 방법과 가중치 감소 기법이 있다.

(6) 모든 입력값과 출력값이 0과 1사이의 범위에서 정의되어야 한다. 입력값이 특정 범위(대개 0과 1 사이)의 값을 가져야 하기 때문에 입력값에 대한 변환 작업이 필요하다.

2.5 평가 방법

모형 평가란, 모형을 구축한 후에 모형의 예측 성능을 확인하고 고려된 서로 다른 모형들 중 어느 모형이 가장 우수한 예측력을 보유하고 있는 지 등을 비교 분석하는 과정으로, 본 논문에서는 Fraud(보험금 부당 청구건)와 Non-Fraud(보험금 정당 청구건) 두 가지 값만을 갖는 이진 종속변수를 분류하는 모형에서 주로 사용되는 모형 평가 방법인 오분류표(confusion matrix), ROC 그래프 그리고 리프트 차트(Lift Chart)에 기반하여 모형을 평가한다.

2.5.1 오분류표

훈련과 검증 과정을 거쳐 구축된 각 모형에 새로운 데이터인 평가용 데이터를 적용해 표 1과 같이 목표 변수의 실제 범주와 모형에 의해 예측된 범주 사이의 관계를 나타내는 오분류표(confusion matrix)를 작성해 예측 모형의 성능을 평가할 수 있다. 오분류표에서 범주 별로 정분류한 빈도는 대각선 원소에, 오분류한 빈도는 비 대각선 원소에 나타나게 된다.

오분류표에서 예측의 정확성과 관련하여 다음과 같은 용어가 사용된다.

표 1: 오분류표

실제 \ 예측	정상	사기	오분류율
정상	TN	FN	실제 0의 도수
사기	FP	TP	실제 1의 도수
	예측 0의 도수	예측 1의 도수	

TN(True-Negative), FP(False-Positive), FN(False-Negative), TP(True-Positive)

1. 정확도(Accuracy, 정분류율) = (TP+TN)/(TP+TN+FP+FN)
2. 오분류율(Missclassification Rate) = (FP+FN)/(TP+TN+FP+FN)
3. 특이도(Specificity, true-negative rate) = TN/(FP+TN)
4. 1-특이도(false-positive rate) = FN/(FP+TN)
5. 민감도(Sensitivity, true-positive rate) = TP/(TP+FN)

2.5.2 ROC 그래프

ROC (Receiver Operation Characteristic) 곡선은 오분류표에서 구한 민감도를 수직축, 1-특이도를 수평축으로 하여 연결선을 그린 것으로 분류 모형의 성능을 평가하기 위해 쓰이는 그래프이다. 예측력이 전혀 없는 모형에 의한 ROC 곡선은 대각선으로 나타나기 때문에 좋은 예측력을 가진 모형일수록 ROC 곡선 아래의 면적(ROC index) 값이 큰 값을 가지게 된다. 따라서 ROC index 는 모형의 성능을 측정하는 값으로 사용된다. 대각선의 밑면적이 0.5 이므로 ROC index 는 최소 0.5에서 최대 1 의 값을 가지게 되며, 1에 가까울수록 예측력이 좋다는 것을 의미한다.

ROC index는 민감도와 '1-특이도'를 각각 SE_i와 SP_i이고 r을 등급의 수라고 할 때, 다음과 같은 식에 의해 구할 수 있다.

$$\text{ROC index} = \sum_{i=2}^{r} (SE_i + SE_{i-1})(SP_i - SP_{i-1})/2$$

2.5.3 리프트 차트

리프트 차트(Lift Chart)는 사후확률에 의해 모형을 평가하는 방법으로 다음과 같이 구해진다.

1. 분석된 모형의 모든 개체에 대한 사후확률을 구한다.
2. 각 모형 별로 사후확률의 내림차순으로 전체 자료를 정렬한다.
3. 정렬이 끝난 자료를 균일하게 K 등분하여 등급화한다.
4. 각 등급에서 목표변수의 특정 범주에 대한 빈도를 구한다.
5. 각 등급에서 다음과 같은 통계량을 계산한다.

이렇게 구해진 차트에서는 다음과 같은 값으로 모형을 평가할 수 있다.

- 반응률(%Response): 각 등급에서 목표범주 1의 비율을 나타낸다.
- 반응검출률(%Captured Response): 각 등급에 목표범주 1에 속하는 개체들이 비율
- 향상도(Lift): 각 등급에서의 반응률이 기준선 반응률에 비해 얼마나 높은지를 나타낸다.

여기서 각 등급은 사후확률에 따라 매겨진 순위이므로 상위 등급에서는 높은 반응률과 향상도, 하위 등급에서는 상위 등급과 차이 나게

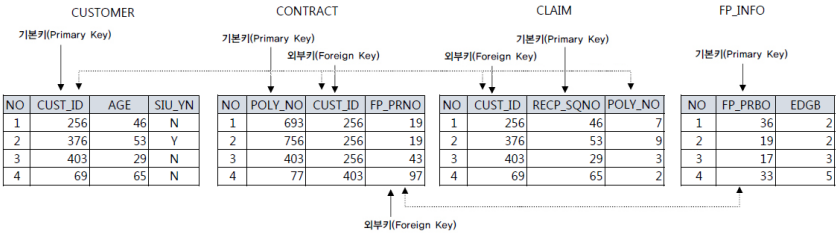
낮은 반응률과 향상도를 보여야 좋은 예측 모형이라 할 수 있다. 만약 등급에 관계 없이 반응률이나 향상도에 별 차이가 없다면 이는 예측 모형의 성능이 좋지 않음을 나타낸다.

제 3 장

분석 데이터

분석에 사용된 데이터는 국내 모 보험회사의 보험가입자 20,607 명이 보유하고 있는 104,014 개의 보험계좌와 109,773 건의 보험청구에 대한 정보 및 보험설계사 31,522 명의 정보를 저장한 4개의 테이블에서 추출하였다. 즉, 보험회사에서 보험과 관련된 데이터베이스는 보험가입자 정보(CUSTOMER), 보험금 청구 관련 정보 (CLAIM), 가입된 보험 관한 정보 (CONTRACT) 및 보험설계사 정보 (FP_INFO)등으로 구분하여 저장된 관계형 데이터베이스이다. 각 테이블에서 필요한 정보는 테이블의 키(key)인 고객 아이디(CUST_ID), 계약별 접수일련 번호(RECP_SQNO), 증권번호(POLY_NO), 보험설계사 사번(FP_PRNO)를 이용하여 참조할 수 있다. 이는 그림 3와 같이 나타낼 수 있다.

CUSTOMER 테이블은 주로 고객의 인구통계적 정보와 신용등급, 최초 고객 등록일, 소득, 총 보험 납입 금액 등의 데이터로 구성되어 있고, CLAIM 테이블에서는 보험금 청구 일자, 사고원인, 입원기간, 치료병원 구분, 실손처리 여부, 청구금액, 지급금액 등을 저장되어 있다. 또, CONTRACT 테이블에서는 고객이 가입한 보험의 종류, 보험상품을 구입한 채널, 보험의 만기일, 주보험금 및 합계보험금 등의 데이터가 있다. 한편 FP_INFO 테이블은 재직 여부, 입사년월, 퇴사년월, 학력 및 설계사 이전의 직업에 관한 정보를 담고 있다.



NO	CUST_ID	AGE	SIU_YN
1	256	46	N
2	376	53	Y
3	403	29	N
4	69	65	N

NO	POLY_NO	CUST_ID	FP_PRNO
1	693	256	19
2	756	256	19
3	403	256	43
4	77	403	97

NO	CUST_ID	RECP_SQNO	POLY_NO
1	256	46	7
2	376	53	9
3	403	29	3
4	69	65	2

NO	FP_PRBO	EDGB
1	36	2
2	19	2
3	17	3
4	33	5

그림 3: 보험사의 관계형 데이터베이스

3.1 보험사기 패턴에 따른 변수의 생성

적절한 데이터의 확보와 데이터로부터 해당 문제를 잘 표현할 수 있는 변수들의 생성이 성공적인 데이터 마이닝의 관건이라고 할 수 있다. 즉, Guha et al. ( )는 적절한 입력변수(input feature)가 예측 모형의 예측력을 향상시킬 뿐 아니라, 문제의 복잡성을 단순화시킬 수 있는 효과가 있다고 하였다. 특성 생성(feature engineering)은 원시데이터로부터 예측에 보다 적합한 변수들의 추출 또는 생성하는 과정으로서 성공적인 데이터 마이닝을 위해서는 반드시 필요한 작업이다. 한편 데이터 마이닝을 필요로 하는 분야에 대한 지식은 특성 생성(feature engineering)이 성공적으로 수행되기 위한 조건이라고 할 수 있다.

보험사기 적발을 위한 데이터 마이닝에서는 4 개의 테이블로 구성된 데이터베이스에서 문제의 해결에 유용한 정보 또는 변수가 적절히 추출될 수 있어야 하며 이를 위해서는 보험과 관련된 지식이 필요하다. Kirlidog and Asuk (2012)에 의하면 보험사기의 경우 문화적인 차이에 따라 나라별로 조금씩 다른 경향이 있지만, 보험전문가에게는 이미 알려져 있는 몇 가지 패턴이 있다고 한다. 그들이 제시한 패턴을 요약하면 다음과 같다.

- 과도한 의료비용
- 짧은 기간(3-4일) 기간안에 많은 보험청구
- 보험청구일자와 보험계약일(만기일)의 차이가 많지 않은 경우
- 많은 횟수의 보험청구
- 과도한 치료기간
- 보험사기와 관련이 많은 병원을 이용

한편 보험사기 경험이 없는 대다수의 가입자는 보험의 필요성 때문에 보험을 가입하기 때문에 보험에 대한 충성도(loyalty)가 높을 것을 예상할 수 있다. 따라서 고객의 충성도를 측정할 수 있는 변수들도 필요하다.

Kirlidog and Asuk (2012)는 앞에서 언급된 패턴에 따르는 보험청구에 대해서는 면밀한 조사가 필요하며, 조사는 보험당사자 뿐 아니라 해당 보험을 담당한 보험 에이전시(insurance agency)에 대한 조사도 필요하다고 하였다. 또, Phua et al. (2010)에서도 보험 에이전시와 보험사기가 직간접 관계가 있음을 기술하였다. 그림 4은 분석에 사용된 데이터에서 보험설계사 별로 사기 보험청구 관련 건수를 조사한 것이다. 이 그림에서 보험사기와 관련된 82명의 보험설계사는 2건 이상의 보험사기와 관련이 있는 것으로 나타났다. 즉, 보험사기와 관련된 2369건의 보험청구 중 1121건(47.3%)은 82명의 보험설계사와 관련이 있었다. 이는 보험설계사와 보험사기의 관련은 일회성이 아닐 수 있다는 것을 의미하기 때문에 보험사기 예측에서 보험설계사에 대한 정보는 매우 중요하다고 할 수 있다.

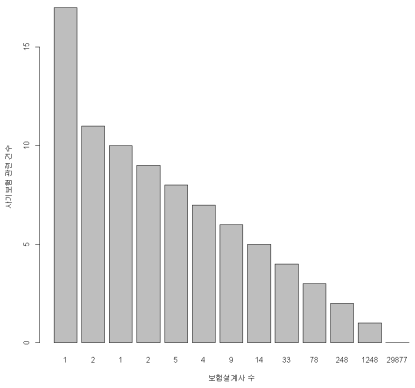


그림 4: 보험설계사별 사기 보험청구 관련 건수

본 연구에서는 보험설계사의 사기 보험청구 관련 건수를 이용하여 담당 보험설계사의 리스크를 측정하는 변수를 생성하였다. 그러나, 이 변수는 타겟변수의 값을 이용하여 구해지기 때문에 입력변수로 사용될 수 없다. 따라서, 포아송 회귀모형과 로짓회귀모형을 이용하여 보험 설계사 리스크의 예측값을 구하였으나, FP_INFO 테이블에는 보험설계사에 대한 충분한 정보가 없었기 때문에 리스크 추정을 위한 모형의 예측오차가 매우 컸으며 예측된 리스크 값은 보험사기 예측모형에서 설명력이 없는 것으로 나타났다.

보험설계사의 리스크는 중요한 입력변수이기 때문에 보험사기 예측이 필요한 보험회사에서는 보험설계사에 대한 보다 많은 정보와 사기 관련 이력을 데이터베이스화 할 필요가 있다. 향후, 설계사의 리스크를 입력변수로 사용할 수 있다면 보험사기 예측 모형의 예측력은 크게 향상될 수 있을 것으로 보인다.

보험사기 패턴을 반영하여 생성된 변수와 변수에 대한 설명은 표 2과 같다. 표에서 CAUS_CODE_YN, DMND_RESN_CODE_YN, HOSP_

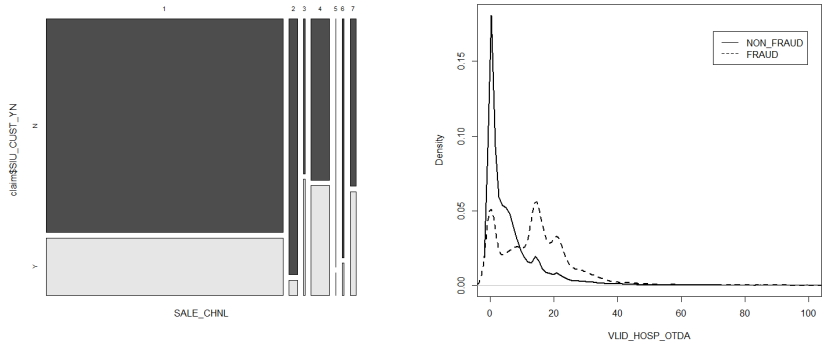
표 2: 보험사기 패턴을 반영하여 생성된 입력변수

변수명	변수 설명
N_OF_INSUR	보험 가입 갯수
N_OF_KIND	보험 종류 갯수
N_OF_CLAIM	보험청구 횟수
MAIN_GOOD	주 보험 종류 코드
CAUS_CODE_YN	사고원인에서 사기 빈도가 높은 코드 이면 “Y”
DMND_RESN_CODE_YN	청구사유가 입원 또는 장애이면 “N”
VALI_HOSP_MAX	보험청구 중에서 입원/통원 최대 일수
HOSP_DVSN_YN	병원 구분에서 요양병원 또는 한방병원이면 “Y”
DIFF_AMOUNT_MAX	보험청구에서 청구액과 지급액 차이의 최대값
DIFF_AMOUNT_MIN	보험청구에서 청구액과 지급액 차이의 최소값
FP_RISK	보험설계사 리스크(추정값)
SALE_CHNL_YN	보험판매 채널에서 설계사, 법인이 아니면 “Y”
PAYM_TERM_YN	보험료 납입 기간이 25개월 미만이면 “Y”
MAX_PAY_DF	최대보험료 납입일과 청구일 간의 기간 차이
CHANG_FP_YN	보험청구 중 설계사가 변경된 청구가 있으면 “Y”

DVSN_YN, SALE_RISK_YN, PAYM_TERM_YN 등은 모형의 예측력을 높히기 위해 그림 5(a)와 같은 모자이크 그림(mosaic plot)을 검토하여 세분화되어 있는 코드를 이분화하여 구성한 변수이다.

3.2 변수 선택

보험사기 패턴에 의한 생성된 변수 외에도 4 개의 테이블에는 보험 및 보험청구에 대한 많은 변수가 있다. 이들 가운데 일부는 사기 적발에 유용할 수 있기 때문에 테이블 상의 변수들을 변수의 특성에 따라 모자이크 그림이나 그림 5(b)와 같은 커널 밀도함수 추정(kernel density etimation)을 통해 보험사기와의 관련성을 검토하여 표 3과 같이 보험사기 적발 모형의 입력변수를 선정하였다. 관련성 판단이 어려웠던 변수들의 변수선택(feature selection)은 분석모형에 의존하도록 하였다.



(a) SALE_CHNL의 모자이크 그림 (b) VLID_HOSP_OTDA의 밀도추정

그림 5: 모자이크 그림과 커널 밀도추정 그림

변수 선정과정에서 특히 CLAIM 테이블의 VLID_HOSP_OTDA는 각 보험청구에서 입원/통원 치료기간에 대한 변수로서 보험사기 패턴의 “과도한 치료기간”과 관련이 있다. 그림 5(b)는 VLID_HOSP_ OTDA에 대해 FRAUD와 NON-FRAUD를 구분하여 밀도함수를 추정한 것으로 FRAUD의 치료 일수는 비교적 큰 경향이 있는 것을 확인할 수 있다. 그러므로 VLID_HOSP_OTDA는 적절한 입력변수라고 판단할 수 있다. 그러나, 이 변수는 각 보험청구에서의 “치료 일수”이므로 한 명의 보험청구인이 여러 개의 치료기간 값이 가질 수 있기 때문에 이를 보험가입자 별 하나의 값으로 묶을 수 있어야 한다. 앞서 살펴 보았듯이 FRAUD의 경우 치료기간이 긴 경향이 있고, 본 논문의 목적이 FRAUD의 적발에 있기 때문에 VLID_HOSP_OTDA의 평균보다는 각 보험청구인의 치료기간 중 최대값이 FRAUD의 적발에 유용하다고 판단된다. 따라서 최대값을 VALI_HOSP_MAX라는 변수로 설정하였다. 이 변수는 표 2에서와 같이 보험 사기 패턴 관련 변수로 분류하였다.

본 연구에서 보험사기 적발 모형의 개발에 사용된 데이터는 목표

표 3: 변수 선택과정에서 선정된 입력변수

변수명	변수 설명
CUST_FRAUD_YN	Target
SEX	성별
AGE	나이
RESI_COST	주택 가격
RESI_TYPE_CODE	거주지 형태
FP_CAREER	보험가입자의 FP 경력 “Y” = yes, “N” = no.
CUST_RGST	최초 고객 등록일 부터 현재까지의 월 수
OCCP_GRP	직업코드 8개
TOTALPREM	납입 총액
MINCRDT	신용등급 최소값
MAXCRDT	신용등급 최대값
WEDD_YN	결혼 여부
MATE_OCCP_GRP	배우자 직업코드
CHLD_CNT	자녀수
LTBN_CHLD_AGE	막내 자녀 나이
MAX_PRM	추정 소득
CUST_INCM	고객 소득
RCBASE_HSHD_INCM	고객 가구 추정소득
RESL_CD1	사고원인 코드
RESL_NM1	사고결과 코드
PMMI_DLNG_YN	실손 처리여부
HEED_HOSP_YN	금감원 유의 병원 여부

변수(target)인 CUST_FRAUD_YN과 36 개 입력변수에 대한 20,607 개의 관찰값이다. 관찰값 중 18,801 개와 1,806 개의 관측값들이 각각 정상과 보험 사기 고객의 데이터로써 2 개 집단의 관찰값 수가 극히 불균형하다. 이와 같은 클래스 분포(class distribution)의 불균형은 데이터 마이닝에서 흔히 나타나는 문제로서, Ezawa et al.(1996), Radivojac et al. (2004) and Chawla (2004)에서 불균형 데이터에 대한 데이터 마이닝 방법을 다루고 있다. 즉, 클래스 분포가 불균형인 데이터에 의한 모형의 추정은 다수의 관찰값이 속한 클래스로 과적합되는 경향이 있기 때문에 보험 사기 적발 모형에서 중요시 되는 참-거짓 오류가 발생할 가능성이 높

아 진다. 그러므로 불균형 데이터에 의한 데이터 마이닝 방법론에 대한 많은 연구(Japkowicz, 2000; Dietterich, 2000; Weiss and Provost, 2003)가 있다.

일반적인 데이터 마이닝 소프트웨어에서 불균형 데이터의 문제는 소위 오버샘플링(oversampling)이라고 불리우는 층화추출법을 사용하여 해결할 수 있다. 즉, 훈련용 데이터(training data)는 클래스 분포가 균형을 이루도록 각 클래스의 표본 수를 조정하여 클래스 별로 임의표본추출하고, 클래스의 불균형을 사전분포(prior distribution)을 이용하여 조정하는 것이다. 본 연구에서는 FRAUD와 NON-FRAUD 클래스에서 각각 1500 개와 2500 개의 표본을 임의 추출하고, 이를 다시 7:3의 비율로 나누어 훈련용 및 분석용 데이터(validation)를 구성하였고, 층화표본추출에서 남겨진 16,107 개의 관찰값은 검증용 데이터(test data)로 사용하기로 한다.

제 4 장

분석 결과

4.1 로지스틱 회귀모형

회귀분석에서 모든 변수를 사용하는 것은 모형의 복잡성을 증가시키므로 변수선택 방법 중 단계별 선택방법(stepwise selection method)를 실시하였고, 16단계에 걸쳐 AGE, CAUS_CODE_YN, CHANG_FP_YN, DMND_RESN_CODE_ YN, FP_CAREER, HEED_HOSP_YN, HOSP_DVSN_YN, MAXCRDT, MAX_PRM, OCCP_GRP, N_OF_CLAIM, PAYM_TERM, PMMI_DLNG_YN, SALE_RISK, VALI_HOSP_MAX가 선택되었다. 선택된 모형의 타당성을 검증하기 위해 가능도비 검정을 실시한 결과, 자유도 47에서 $\chi^2 = 1572.1887$으로 나타나 p-val $< .0001$이므로 적절한 모형이라고 판단된다. 적합된 모형의 결과와 선택된 변수들 중 유의한 변수들의 회귀계수 추정치는 표 4과 같다.

로지스틱 회귀 분석의 오즈비(odds ratio) 인 Exp(Est) 추정값으로 각 변수의 영향력을 해석하면 다음과 같다.

- CAUS_CODE_YN: 사고원인의 빈도분석을 통해 상대적으로 보험사기가 적은 것으로 분류된 코드(변수의 값이 "N")이면 그렇지 않은 코드보다 보험 사기일 확률이 0.524배로 감소.
- CHANG_FP_YN: 보험 청구 당시 보험관리인이 보험계약시의 보험설계사와 같으면 같지 않은 경우와 비교하여 보험 사기일 확

표 4: 로지스틱 회귀모형 추정결과

-2 Log Likelihood		Likelihood Ratio		
Intercept only	Intercept & Covariate	Chi-square	DF	Pr > ChiSq
3708.816	2136.628	1572.1887	47	< .0001

변수		Estimate	Standard Error	Wald Chi-square	p-val	Exp(est)
AGE		-0.0132	0.00493	7.16	0.0075	0.987
CAUS_CODE_YN	N	-0.647	0.0779	68.93	< .0001	0.524
CHANG_FP_YN	N	-0.169	0.059	8.22	0.0041	0.844
DMND_RESN_CODE_YN	N	0.6842	0.0661	107.14	< .0001	1.982
FP_CAREER	N	-0.4342	0.1148	14.32	0.0002	0.648
HEED_HOSP_YN	N	-0.2416	0.1076	5.04	0.0247	0.785
HOSP_DVSN_YN	N	-0.4363	0.0869	25.18	< .0001	0.646
IMP_MAX_PRM		1.85E-07	6.59E-8	7.88	0.005	1.000
IMP_OCCP_GRP	주부	0.5037	0.1321	14.55	0.0001	1.655
IMP_OCCP_GRP	자영업	0.3939	0.1603	6.04	0.014	1.483
IMP_OCCP_GRP	제조업	-0.4334	0.2411	3.23	0.0722	0.648
N_OF_CLAIM		0.1529	0.0119	164.94	< .0001	1.165
PAYM_TERM_YN	N	0.4683	0.1112	17.74	< .0001	1.597
PMMI_DLNG_YN	N	0.5205	0.0858	36.82	< .0001	1.683
SALE_CHNL_YN	N	-0.4145	0.0847	23.93	< .0001	0.661
VALI_HOSP_MAX		0.0201	0.00302	44.24	< .0001	1.020

표 5: 로지스틱 회귀모형의 오분류표

분석용				검증용			
예측 / 실제	정상	사기	오분류율	예측 / 실제	정상	사기	오분류율
정상	584	56	8.75%	정상	125	13	9.42%
사기	124	268	31.63%	사기	23	72	24.21%
전체 오분류율 17.44%				전체 오분류율 15.45%			

률이 0.844배로 감소.

- DMND_RESN_CODE_YN: 청구사유가 입원 또는 장애이면 사망, 통원 등의 다른 사유보다 1.982배로 증가.

- HEED_HOSP_YN: 금감원 유의 병원이 아닌 병원에서 치료하였

으면 유의 병원에서 치료한 경우와 비교하여 보험 사기일 확률이 0.785배로 감소.

- N_OF_CLAIM: 청구 횟수가 1회 증가할수록 보험 사기 확률이 1.165배로 증가.
- PMMI_DLNG_YN: 실손 처리를 하지 않으면 실손 처리를 하는 것에 비해 보험 사기 확률이 1.683배로 증가.
- SALE_CHNL_YN: 보험구입을 설계사 또는 법인을 통해 구입하였으면 인터넷, 전화, 방카슈랑스 등의 다른 구매 채널에 비해 보험 사기 확률이 0.661배로 감소
- VALI_HOSP_MAX: 최대 유효 입원 일수가 1일 증가할수록 보험 사기 확률이 1.02배 증가한다.
- OCCP_GRP: 보험청구인의 직업이 주부 또는 자영업이면 다른 직업군과 비교하여 각각 1.655, 1.483배 증가하고, 제조업이면 0.648배로 감소

이와 같은 보험사기 확률의 증가 또는 감소의 해석은 보험사기 패턴과 일반적인 상식에 부합되므로 추정결과를 신뢰할 수 있다고 판단된다.

추정된 로지스틱모형에 검증용 데이터와 평가용 데이터를 적용하여 구한 오분류표(confusion matrix)는 표 5과 같다. 분석용 데이터로 추정된 모형을 검증용 데이터에 적용했을 때, 정상 거래 고객을 보험 사기자로 예측하는 즉, 거짓-음(false-negative)오류의 확률은 13/138 = 0.0942이고, 보험사기자를 정상 고객이라고 예측하는, 즉 거짓-양(false-positive)의 오류 가능성은 0.2421이다. 보험사기 예측 모형에서는 거짓-양(false-positive)의 오류확률이 최소가 되도록 하여야 한다.

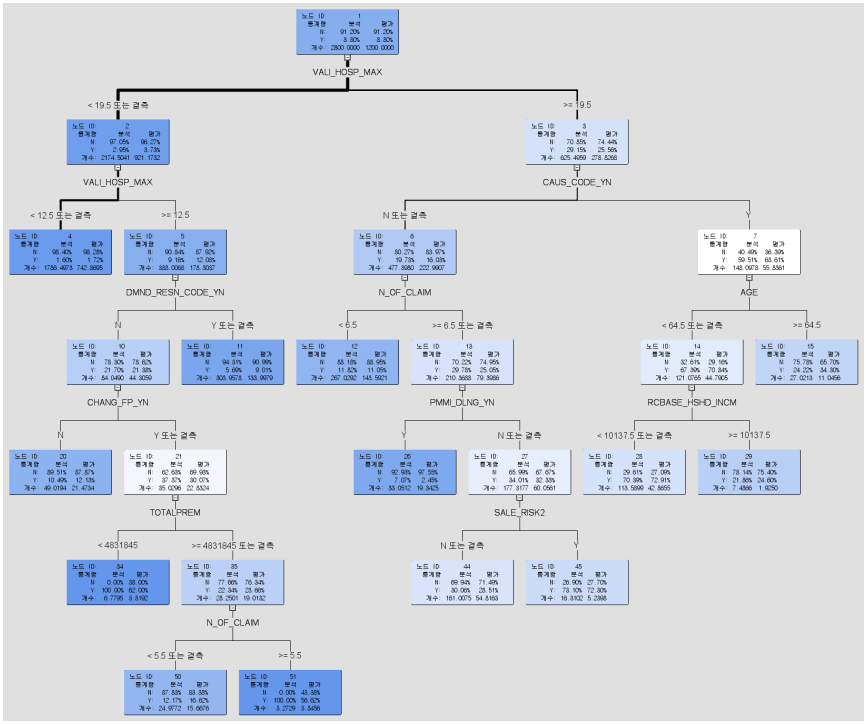


그림 6: 보험사기 적발 의사결정나무

4.2 의사결정나무

세 가지 분리 기준에 따라 각기 구축된 의사결정나무 중에서 ROC, 오분류율 등을 비교하여 Gini 계수에 의해 분리된 나무가 최종 선정되었다. 선정된 나무는 그림 6과 같이 13 개의 터미날 노드(terminal node)를 갖고 있다. 이 모형에서 보험 사기 여부에 대한 첫 번째 분리 기준은 VALI_HOSP_MAX 임을 확인할 수 있고, 19.5일을 기준으로 최대 유효 입원 일수가 많으면 보험 사기일 확률이 높게 나타난다. 의사결정나무 모형을 구축하는 데에 중요한 변수와 변수의 중요도는 표 6와 같고, 이를 시각화한 것이 그림 7이다.

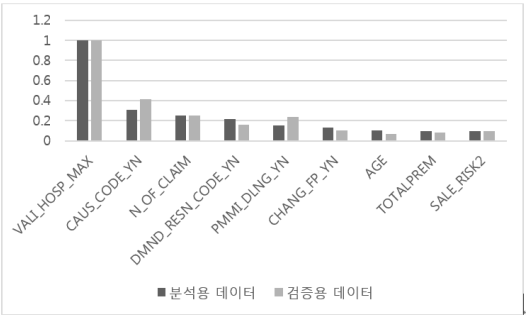


그림 7: 의사결정나무에서 변수의 중요도

표 6: 의사결정나무에서 변수의 중요도

변수이름	분석용 데이터 중요도	평가용 데이터 중요도
VALI_HOSP_MAX	1	1
CAUS_CODE_YN	0.309208	0.411125437
N_OF_CLAIM	0.249829	0.250384058
DMND_RESN_CODE_YN	0.215899	0.161924852
PMMI_DLNG_YN	0.153188	0.238491201
CHANG_FP_YN	0.130669	0.102721548
AGE	0.104849	0.068095362
TOTALPREM	0.096485	0.0835044
SALE_CHNL_YN	0.094713	0.096158541

변수 중요도는 구축된 의사결정모형에서 해당 입력 변수가 제외되었을 때 예측력이 얼마나 변하는지를 상대적 크기로 나타낸 것으로 VALI_HOSP_MAX 변수가 제거되면 의사결정나무 자체가 없어지므로 중요도가 1로 계산된다(강현철 외, 2014). 변수 중요도는 분석용 데이터에서의 중요도와 평가용 데이터에서의 중요도 모두 나타나고 분석용 데이터와 평가용 데이터에서 크게 차이 나지 않는 것을 볼 수 있다.

표 7은 분석용 및 평가용 데이터에 대한 의사결정나무의 오분류표이다. 의사결정나무 모형으로 구해진 오분류율(misclassification rate)은

표 7: 의사결정나무의 오분류표

분석용				검증용			
실제 \ 예측	정상	사기	오분류율	실제 \ 예측	정상	사기	오분류율
정상	1,590	176	9.96%	정상	316	41	11.48%
사기	316	718	30.56%	사기	65	178	26.75%
전체 오분류율 17.57%				전체 오분류율 17.67%			

분석용 데이터와 검증용 데이터에서 크게 차이 나지 않음을 볼 수 있다. 또, 분석용 데이터와 검증용 데이터에서 모두 사기를 예측하는 참-양(true-positive)의 가능성이 거짓-양(false-positive)의 가능성보다 크게 나타나고 있다.

4.3 부스팅

부스팅은 모수와의 오차를 가장 작게 만들기 위해 반복을 통해 그림 8과 같이 평균제곱오차(MSE) 값이 가장 작은 부스팅 계열의 항의 개수를 찾는다. 이렇게 구하여진 부스팅 모형에서 선택된 변수의 중요도는 표 8와 같다.

다음은 분석용 데이터와 평가용 데이터에 부스팅 모형을 적용했을 때의 분류 테이블이다. 부스팅 모형으로 구해진 오분류율(misclassification rate)은 분석용 데이터에서 더 좋은 결과를 보여준다. 앞선 모형들과 마찬가지로 분석용 데이터와 검증용 데이터에서 모두 사기를 예측하는 거짓-양의 가능성이 거짓-음의 가능성보다 크게 나타나고 있지만 앞선 모형들에 비해 거짓-양의 확률이 가장 낮다.

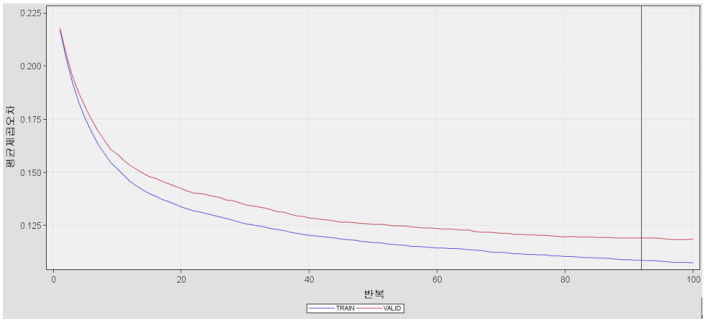


그림 8: 부스팅의 평균제곱오차

표 8: 부스팅 모형에서 선택된 변수의 중요도

변수이름	분석용 데이터 중요도	평가용 데이터 중요도
VALI_HOSP_MAX	1	1
N_OF_CLAIM	0.453345	0.525055
CAUS_CODE_YN	0.330748	0.361106
DMND_RESN_CODE_YN	0.271725	0.360708
MAIN_GOOD	0.205712	0.131251
CUST_INCM	0.193228	0.132227
TOTALPREM	0.175514	0.148478
OCCP_GRP	0.163813	0.110567
MAX_PRM	0.140665	0.105481
HOSP_DVSN_YN	0.134979	0.173522
MAXCRDT	0.12736	0.085825
MATE_OCCP_GRP	0.119447	0.016009
DIFF_AMOUNT_MAX	0.11684	0.105434
RESL_CD1	0.113758	0.118877
PMMI_DLNG_YN	0.101766	0.130454

표 9: 부스팅 오분류표

분석용				검증용			
예측 / 실제	정상	사기	오분류율	예측 / 실제	정상	사기	오분류율
정상	1,651	115	6.51%	정상	317	40	11.20%
사기	207	827	20.02%	사기	50	193	20.58%
전체 오분류율 11.5%				전체 오분류율 15%			

4.4 신경망 모형

신경망 모형을 구축하기 위해 반복적 최적화에 따른 평균제곱오차(average squared error)의 변화를 보면 분석용 데이터에 대한 오차함수 값은 반복횟수가 늘어남에 따라 감소하지만 평가용 데이터에 대한 오차함수 값은 어느 정도 감소하다가 다시 증가하고 있다. 이때, 평가용 데이터의 오차함수 값이 최소가 되는 반복에서의 추정치를 선택하는데 그림 9에서와 같이 반복횟수 2에서 분석용 데이터의 오차함수의 값이 최소가 됨을 확인할 수 있다.

표 10은 은닉층이 1개, 은닉마디가 3개인 신경망 모형을 적용했을 때의 분류 테이블이다. 신경망 모형으로 구해진 오분류율은 분석용 데이터에서 더 좋은 결과를 보여준다. 앞선 모형들과 마찬가지로 분석용 데이터와 검증용 데이터에서 모두 사기를 예측하는 거짓-양의 가능성이 거짓-음의 가능성보다 크게 나타나고 있다.

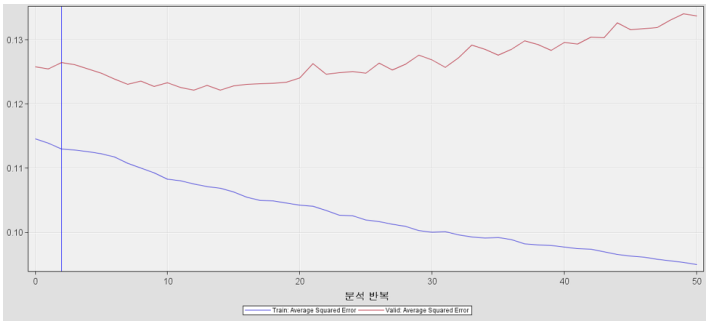


그림 9: 신경망 모형의 평균제곱오차

표 10: 신경망 모형의 오분류표

분석용				검증용			
실제 \ 예측	정상	사기	오분류율	실제 \ 예측	정상	사기	오분류율
정상	599	41	6.41%	정상	125	13	9.42%
사기	111	281	28.32%	사기	32	63	33.68%
전체 오분류율 14.73%				전체 오분류율 19.31%			

제 5 장

모형 평가

모형 평가 과정은 모형을 구축한 후에 모형의 예측 성능을 확인하고 고려된 서로 다른 모형들 중 어느 것이 가장 우수한 예측력을 보유하고 있는 지 등을 비교 분석하는 과정이다. 로지스틱 회귀모형, 의사결정나무 모형, 부스팅 모형, 신경망모형을 적합시킨 후 오분류율, 리프트 차트, ROC 그래프로 모형을 평가하고 최종 모형을 선택하고자 하였다.

5.1 오분류표

보험 사기자를 예측하는 것이 목표이기 때문에 특이도(Specificity, true-negative rate)를 관심을 가지고 봐야 한다. 평가용 데이터에 대한 4 개의 지도학습 모형의 오분류표(표 11)에서 부스팅 모형의 1-특이도(false-positive rate) 값이 23.86%로 가장 작게 나타났으므로 특이도값이 가장 크고 보험 사기자를 예측하는 성능이 가장 높다고 판단할 수 있다. 한편 전체 오분류율은 로지스틱 회귀모형에서 가장 작게 나타났지만 부스팅과 0.05% 정도의 차이만 가지고 있어 전반적으로 부스팅 모형이 선호된다.

표 11에서 각 모형의 표본 개수가 다른 이유는 결측값 때문이다. 의사결정나무와 부스팅은 모형 추정과정에서 결측값을 대체하는 알고리즘이 있기 때문에 평가용 데이터 모두에 대한 예측 결과를 제시하고

표 11: 추정된 4 개 지도학습 모형의 평가용 데이터 오분류표

로지스틱 회귀분석			
실제 \ 예측	정상	사기	오분류율
정상	5,285	703	11.74%
사기	34	83	29.06%
전체 오분류율 12.07%			

의사결정나무			
실제 \ 예측	정상	사기	오분류율
정상	14,322	1,979	12.14%
사기	98	208	32.03%
전체 오분류율 12.51%			

신경망 모형			
실제 \ 예측	정상	사기	오분류율
정상	3,107	436	12.31%
사기	24	43	33.33%
전체 오분류율 12.72%			

부스팅			
실제 \ 예측	정상	사기	오분류율
정상	14,360	1,941	11.91%
사기	73	233	23.86%
전체 오분류율 12.13%			

있으나, 로지스틱 회귀모형이나 신경망 모형의 경우 입력변수가 결측이면 예측값을 구하지 못한다. 입력변수에 결측값을 대체(imputation)하여 예측값을 구할 수도 있으나, 이 경우 모형의 예측력이 왜곡될 수 있고, 본 논문의 연구 목적이 보험사기 적발 모형의 구축보다는 모형의 비교에 관심이 있기 때문에 대체방법을 적용하지 않았다.

5.2 ROC 그래프

그림 10의 ROC 그래프에서는 부스팅 모형이 대각선에서 가장 멀리 나타나는 것을 볼 수 있다. 즉, ROC 그래프에서는 ROC 곡선 아래의 면적(ROC index)으로 모형을 평가하는 바, 이를 수치로 나타내면 표 12와 같다.

5.3 리프트 차트

그림 11의 리프트 차트에서도 큰 차이를 보이지 않지만 로지스틱 회귀 모형이 근소하게 가장 좋은 향상도를 보이는 것으로 판단된다.

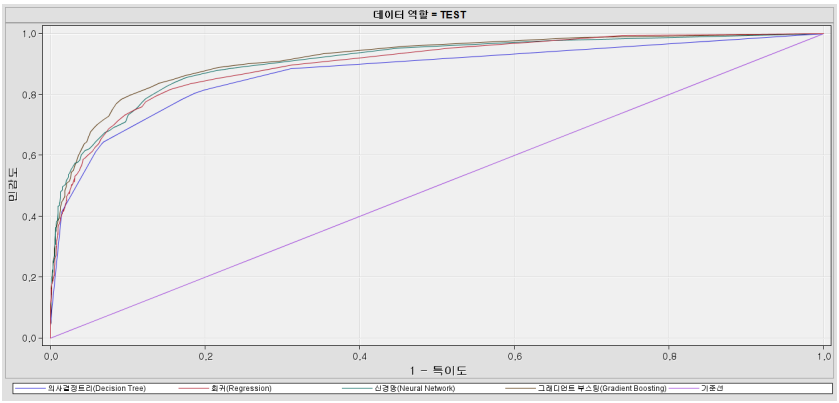


그림 10: 평가용 데이터에 대한 4 개 지도학습 모형의 ROC 그래프

표 12: 평가용 데이터에 대한 4 개 지도학습 모형의 ROC Index

	지도학습 모형			
	로지스틱	의사결정나무	부스팅	신경망
ROC Index	0.9	0.872	0.918	0.909

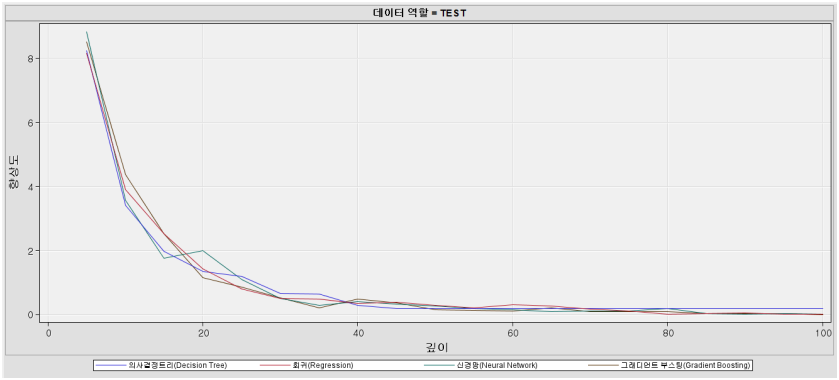


그림 11: 평가용 데이터에 대한 4 개 지도학습 모형의 리프트 차트

종합 요약

이상의 평가를 결과를 요약하면 다음과 같다.

	지도학습 모형			
	로지스틱	의사결정나무	부스팅	신경망
오분류율	12.07%	12.51%	12.13%	12.72%
특이도	0.7094	0.6797	0.7614	0.6667
ROC Index	0.9	0.872	0.918	0.909
ASE	0.1189	0.1274	0.1076	0.1130

제 6 장

결론

전체 모형의 오분류율이 가장 낮은 모형은 로지스틱 회귀모형이다. 하지만 보험 사기 예측에서 주목해야 할 것은 보험 사기자를 옳게 예측하는 즉, 참-거짓 비율이 중요하므로 특이도 값이 모형 평가의 기준이 될 필요가 있다. 특이도를 보면 부스팅 모형이 다른 모형들에 비해 월등히 높은 성능을 보여주며 전체 오분류율도 로지스틱 회귀 모형과 크게 차이 나지 않는 것을 확인할 수 있다. ROC index 값도 부스팅 모형에서 가장 큰 값을 보여주고 ASE(Average Squared Error, 평균제곱오차)값 또한 부스팅 모형이 가장 작은 값을 가지기 때문에 부스팅 모형의 성능이 가장 높다고 평가할 수 있다. 하지만 부스팅 모형의 단점으로 설명력이 부족하고 직관적이지 않다는 점에서 로지스틱 회귀 모형도 보험사기 적발 모형으로 유용할 수 있다고 판단된다.

4개의 모형에서 공통적으로 꼽은 중요 변수는 CAUS_CODE_YN, DMND_RESN_CODE_YN, N_OF_CLAIM, PMMI_DLNG_YN, VALI_HOSP_MAX 로서 실손처리 여부에 대한 이진변수인 PMMI_DLNG_YN를 제외하면 특성 생성(feature engineering)과정에서 생성된 변수이다. 즉, 보험사기 적발 모형에서 보험사기 패턴에 대한 정보가 매우 중요하다는 것을 알 수 있다. 특히, 변수에 대한 검토에서 살펴 보았듯이 보험사기는 보험설계사와 밀접한 관계가 있기 때문에 보다 효율적인 보험사기 적발 모형이 구축되려면 보험설계사의 리스크를 예측할 수 있는

데이터베이스 또는 보험설계사의 사기 관련 이력에 대한 데이터가 확보되어야 하며 이 경우, 본 논문에서 제시한 보험사기 적발 모형은 예측의 정확성을 더욱 높힐 수 있다고 판단된다.

참고 문헌

[1] 강현철, 한상태, 최종후, 이성건, 김은석, 엄익현 (2014). "빅데이터 분석을 위한 데이터 마이닝 방법론," 자유아카데미.

[2] 금융감독원 (2016). "금융감독원 2015년도 보험사기 적발통계," http://www.fss.or.kr

[3] 송윤아 (2010). "보험사기 영향요인과 방지방안", 보험연구원 정책보고서 2010-1.

[4] 송윤아 정인영 (2011). "사기성클레임에 대한 최적조사방안," 보험연구원 정책보고서 2011-5.

[5] 박창이, 김용대, 김진석, 송종우, 최호식 (2013). "R 을 이용한 데이터마이닝," 교우사.

[6] Bologa, A.-R., Bologa, R., and Florea, A. (2010). Big data and specific analysis methods for insurance fraud detection, *Database Systems Journal,* **1**, pp. 30–39.

[7] Chawla, N. V., Japkowicz, N., and Kokz, A., Editors (2004). SIGKDD Special Issue on Learning from Imbalanced Datasets.

[8] Dietterich, T. (2000). An empirical comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization, *Machine Learning,* **40**, pp. 139–157.

[9] Ezawa, K., J., Singh, M., and Norton, S., W. (1996). Learning goal oriented Bayesian networks for telecommunications risk management. *In Proceedings of the International Conference on Machine Learning, ICML-96,* pp. 139-147, Bari, Italy. Morgan Kauffman.

[10] Freidman, J.(2001). Greedy function approximation: A gradient boosting machine, *Annals of Statistics,* **29**, pp. 1189-1232.

[11] Guha, R., Manjunath. S., and Palepu, K. (?). Comprarative analysis of machine learning techniques for detecting insurance claims fraus, http://www.wipro.com/documents/comparative-analysis-of-machine - learning-techniques-for-detecting-insurance-claims-fraud.pdf

[12] Hassan, A., and Abraham, A. (2013). Computational intelligence models for insurance fraud detection: a review of a decade of research, *Journal of Network and Innovative Computing,* **1**, pp. 341–247.

[13] Hastie, T., Tibshirani, R., and Freidman, J. (2009). "The element of statistical learning: data mining, inferenct, and prediction," 2nd Edition, Springer, New York.

[14] Japkowicz, N. (2000b). Learning from imbalanced data sets: A comparison of various strategies. *In Proceedings of the AAAF2000 Workshop on Learning from Imbalanced Data Sets,* Austin, TX.

[15] Kirlidog, M., and Asuk, C. (2012). A fraud detection approach with data mining in health insurance, *Procedia - Social and Behavioral Sciences,* **62**, 989–994.

[16] Phua, C., Lee, V., Smith, K., and Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research, arXiv preprint arXiv:1009.6119.

[17] Radivojac, P., Chawla, N. V., Dunker, K., and Obradovic, Z. (2004). Classification and knowledge discovery in protein databases, *Journal of Biomedical Informatics,* **37** pp. 224-239.

[18] Weiss, G. and Provost, F. (2003). Learning when training data are costly: The effect of class distribution on tree induction, *Journal of Artificial Intelligence Research,* **19**, pp. 315–354.

Abstract

Fraud Detection in Health Insurance Using Data Mining Techniques

MinAh Lee
Department of Statistics
The Graduate School
Seoul National University

The purpose of this study is to predict health insurance fraud which has been a rising issue recently based on statistical model. Logistic regression model and Data Mining techniques such as decision tree, neural network, and boosting are applied to data and performance are compared. To improve the performance, new variables that can represent relying problem the best way possible are constructed through feature engineering the data. The data provided includes customer data, insurance claim data, insurance contract data, and insurance planner data.

The goal of this study to find the most efficient model to predict insurance fraud. Misclassification rate, ROC graph and Lift Chart were used to compare performance and prediction capability among 4 models. Final selected model was boosting model which showed total misclassification rate

of 12.13%. 4 out of 5 variables commonly selected from all 4 models were variables constructed from feature engineering. Those variables can be used to interpret current situation of insurance fraud.

Keywords : Data Mining, Insurance Fraud, Supervised Learning

Student Number : 2015-20306