



## 저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

碩士學位 論文

指導教授 地 元 哲

**Data Mining 기법을 활용한 효율적  
보험사기 적발 모형 연구**

**A Study on Effective Insurance Fraud Detection Model  
using Data Mining Techniques**

弘益大學校 大學院

情報産業工學科

宋 榮 美

2009年 12月 28日



**HONGIK UNIVERSITY**

# Data Mining 기법을 활용한 효율적 보험사기 적발 모형 연구

**A Study on Effective Insurance Fraud Detection Model  
using Data Mining Techniques**

이 論文을 碩士學位 論文으로 提出함.

2009年 12月 28日

弘益大學校 大學院

情報産業工學科

宋 榮 美



**HONGIK UNIVERSITY**

宋榮美의 碩士學位 論文을 推薦함.

2009年 12月 28日

指導教授 地 元 哲

弘益大學校 大學院



宋榮美의 碩士學位 論文을 認准함.

審 查 委 員

審査委員長      朴 丘 鉉 (印)

審査委員      玉 昌 秀 (印)

審査委員      地 元 哲 (印)

弘益大學校 大學院



**HONGIK UNIVERSITY**

## 국문요약

송 영 미

홍익대학교

산업공학과

보험사기란 보험가입자가 부당한 방법으로 거액의 보험금을 편취하는 것으로 최근 경제 불황이 계속되면서 보험사기 비율이 높아지고 있다. 보도 자료에 따르면 보험사기 적발 금액과 혐의자가 최근 3년간 2배로 급증한 것으로 나타났다. 급속한 정보화에 따라 범죄 수법 또한 점차 지능화되고 있다. 지능화되는 보험사기 적발을 위해서는 시스템적 방법의 도입이 필요하다.

본 논문에서는 수집된 데이터를 이용해 그 안에서 통계적으로 의미 있는 규칙이나 패턴들을 찾아내는 다양한 데이터 마이닝 기법들을 적용하여보고 보험사기를 적발 할 수 있는 최적의 방안을 연구하고자 한다. Neural Network, Logistic Regression 등 다양한 데이터마이닝 기법들을 적용해보고 앙상블 모형을 구성하여 보험사기 적발 모형이 가장 효율적인 성능을 낼 수 있는 방안을 모색해 보고자 한다.

모형의 스코어를 각 변수의 해당 범주별 점수로 스케일링하여 만들어진 스코어카드는 보험사기 뿐만 아니라 다른 모형 개발에서도 앙상블 모형이 가진 약점인 해석의 어려움을 해소하는데 도움을 줄 수 있을 것으로 기대된다.



## 차 례

국문요약 .....	i
제목차례 .....	ii
그림차례 .....	iv
표차례 .....	v
제 1 장 서론 .....	1
1.1 연구배경 .....	1
1.2 연구목적 .....	3
1.3 연구방법 .....	4
1.4 연구범위 .....	5
제 2 장 보험사기 현황 .....	6
2.1 보험사기 현황 .....	6
2.1.1 해외현황 .....	6
2.1.2 국내현황 .....	8
2.2 보험사기 유형 .....	12
2.3 보험사기 사례 .....	13
제 3 장 데이터마이닝 기법의 적용 .....	15
3.1 데이터마이닝 기법 .....	15



3.1.1 Decision Tree .....	15
3.1.2 Logistic Regression .....	16
3.1.3 Neural Network .....	16
3.1.4 Ensemble .....	18
3.1.5 Link Analysis .....	18
3.2 스코어카드 생성 .....	20
 제 4 장 앙상블 효과 및 스코어카드 생성 .....	 23
4.1 실험주제 .....	23
4.2 실험개요 .....	24
4.2.1 데이터이해 .....	24
4.2.2 데이터준비 .....	26
4.2.3 모형생성 .....	31
4.3 실험결과 .....	33
4.3.1 성능평가 .....	33
4.3.2 스코어카드 비교 .....	38
 제 5장 결론 및 향후 연구과제 .....	 44
 참고문헌 .....	 46
영문요약 .....	47





## 그림차례

그림 2-1 보험사기 추정 금액 현황 .....	10
그림 3-1 스코어 Scale 조정 범위 설정 .....	20
그림 3-2 스코어카드 생성 예 .....	22
그림 4-1 학습용 샘플 데이터의 구성 .....	27
그림 4-2 대인모형 샘플 데이터 대표성 검증 .....	28
그림 4-3 대물모형 샘플 데이터 대표성 검증 .....	29
그림 4-4 자차모형 샘플 데이터 대표성 검증 .....	29
그림 4-5 자손모형 샘플 데이터 대표성 검증 .....	30
그림 4-6 앙상블 모형의 스코어카드 생성 .....	32



## 표차례

표 2-1 보험사기 적발 현황 .....	8
표 2-2 연령대별 보험사기 적발 인원 현황 .....	9
표 2-3 보험사기 추정 금액 현황 .....	10
표 2-4 손해보험 종류별 보험사기 적발 금액 현황 .....	11
표 2-5 손해보험 종류별 보험사기 적발 인원 현황 .....	11
표 4-1 학습용 샘플 데이터 생성 .....	27
표 4-2 대인 모형 스코어의 %Response 비교 .....	34
표 4-3 대인 모형 스코어의 %Captured Response 비교 .....	34
표 4-4 대물 모형 스코어의 %Response 비교 .....	34
표 4-5 대물 모형 스코어의 %Captured Response 비교 .....	34
표 4-6 자차 모형 스코어의 %Response 비교 .....	34
표 4-7 자차 모형 스코어의 %Captured Response 비교 .....	34
표 4-8 자손 모형 스코어의 %Response 비교 .....	34
표 4-9 자손 모형 스코어의 %Captured Response 비교 .....	34
표 4-10 대인 모형 스코어카드 비교(1) .....	38
표 4-11 대인 모형 스코어카드 비교(2) .....	39
표 4-12 대인 모형 스코어카드 비교(3) .....	40
표 4-13 대물 모형 스코어카드 비교 .....	41
표 4-14 자차 모형 스코어카드 비교 .....	42
표 4-15 자손 모형 스코어카드 비교 .....	43



## 제 1 장 서론

보험사기란 보험가입자가 부당한 방법으로 보험자로부터 거액의 보험금을 편취하는 것을 말한다. 최근 경제 불황이 계속되면서 보험사기 비율이 높아지고 있다. 보도 자료에 따르면 보험사기 적발 금액과 혐의자가 최근 3년간 2배로 급증한 것으로 나타났다. 보험사기의 증가로 보험회사에서는 보험사기 적발 업무의 중요성이 증가하고 있다.

### 1.1 연구배경

금융감독원에서 분석한 2009년도 상반기 보험사기 적발현황에 따르면 보험사기 적발실적은 1,460억 원(22,801명)으로 전년 동기 대비 금액기준 33.6%(367억 원), 인원기준 44.0%(6,969명)로 급증 하였다. 보험사기 적발이 최근 급격히 증가한 원인은 경기침체로 인하여 무직이나 일용직 등 소득기반이 취약한 계층의 생계형 보험사기가 큰 폭으로 증가한 것을 들 수 있다. 또한 금융범죄 근절을 위해 금융감독원과 경찰청간 업무협약을 체결('09. 5. 4)하는 등 보험사기 예방을 위한 특별 단속을 강력하게 추진하게 된 영향도 있다. 이처럼 최근 보험사기 증가에 따라 사기 적발에 대한 관심도 더욱 증가하고 있다.

보험사기는 외견상 보험사에만 직접적인 손해를 입히는 것 같지만 보험사의 보험금 지급액이 증가하면 가입자들이 내야하는 보험료도 증가할 수 밖에 없으므로, 결국 보험사기는 보험사의 경제적 손실뿐만 아니라 선의의 보



험계약자가 더 많은 보험금을 내야하는 피해를 주게 된다. 일부 보험 범죄는 경제적 피해 뿐 만 아니라, 신체에 해를 가하거나 생명을 위협하는 강력 범죄로 이어지고 있어 사회 불안 요인으로 작용하게 된다.

2009년 상반기 보험사기로 적발된 혐의자들을 연령대별로 살펴보면, 10대 및 20대의 증가율은 전년 동기 대비 각각 157.1%, 64.3%로서 다른 연령대(30~60대)의 증가율을 크게 상회하고 있다. 특히, 10대의 보험사기는 과거에는 유흥비 마련을 위해 단순 가담하는 형태였으나, 최근에는 학교 선후배 등과 공모하여 조직적으로 보험사기를 실행하는 형태로 발전하였다. 이처럼 최근에는 청소년이 가담되는 경우도 늘고 있어 심각한 사회 문제로 대두되고 있다.



## 1.2 연구목적

자동차 보험의 대인배상 손해에 관한 실증분석에서 자동차 상해 보험사기의 10%를 차지하는 경성사기의 1%, 50%를 차지하는 연성사기의 2%만이 보험금 지급이 중단되고 있다고 한다(Weisberg & Derrig, 1991). 전체 발생한 보험사기 건의 약 98%는 적발되지 않고 보험금이 지급되어 보험회사의 손해 및 선의의 보험가입자들에게 피해를 주고 있다. 효율적인 보험금 지급과 부당한 손실을 방지하기 위해서는 보험사기 적발 업무의 중요성은 커지고 있다.

전통적인 보험사기 적발 업무는 조사요원들의 개인적인 경험이나 안목 등 인적 능력을 중요시하는 방법이었다. 이러한 방법은 충분한 전문 인력을 확보하기도 어려울 뿐 아니라 원활한 조사 능력을 갖추기까지 많은 비용 및 자원의 투자를 필요로 한다. 이러한 보수적인 방법으로는 사기로 의심이 가는 건에 대해서 조사 인력의 부족과 근거자료 확보의 어려움 등으로 급증하는 보험사기 청구 건을 감당하기에는 한계가 있게 된다.

급속한 정보화에 따라 보험사기 수법이 점차 지능화 되고 있는 상황에서 조사자들의 눈에 보이는 것만 조사해서는 많은 경우 적발되지 못하게 된다. 고도로 지능화되는 사기를 적발하기 위해서는 과학적 보험사기 적발 업무가 이루어져야 한다. 데이터마이닝 기술을 이용하면 대용량 데이터를 실시간으로 파악하고, 복잡한 컴퓨터 알고리즘을 이용해 알아내기 힘든 변화나 새로운 패턴을 쉽게 찾아낼 수 있다.

따라서 본 연구에서는 다양한 데이터마이닝 기법들을 적용해보고 앙상블 모형을 구성하여 보험사기 적발 모형이 가장 효율적인 성능을 낼 수 있는 방안을 모색해 보고자 한다.



### 1.3 연구방법

보험사기 적발 스코어 모델을 생성할 때는 데이터마이닝 분석 기법 중 각 변수의 구간대별 스코어카드 생성이 가능한 Logistic Regression 기법이나 모델의 결과로 규칙형태의 도식화가 가능한 Decision Tree 기법을 사용한다. 이런 기법들로 생성된 모델의 효과성을 비교 검토하여 최종 모델을 결정하여 사용하는 일반적인 방식이다.

본 논문에서는 일반적으로 보험사기 적발모형에 사용하는 Logistic Regression과 앙상블 모형의 결과를 비교하여 앙상블 모형의 우수한 성능을 검증하였다. 앙상블 모형의 성능이 어느 단일모형의 성능보다도 우수하고 안정적이라는 것은 이미 많은 연구에서 증명되었다. 그러나 앙상블 모형이 실제 업무에 적용되기에는 생성된 스코어에 대한 해석이 어렵다는 단점이 있다. 따라서 본 논문에서는 완성된 앙상블 모형에 해석의 용이성을 더하기 위하여 각 변수별 영향력을 파악하는데 도움이 될 수 있는 스코어 카드를 제시하였다.

앙상블 모형의 스코어로 새로운 목표변수를 생성하여 스코어카드 작성이 가능한 기법인 Logistic Regression으로 재학습하는 방법으로 스코어 카드를 생성하였다. 이렇게 생성된 앙상블 모형의 스코어카드는 해석이 어려운 앙상블 모형의 스코어가 실무에 적용 되었을 때 조사요원들이 해당 청구 건에 부여된 스코어를 쉽게 이해하는 데 도움이 될 것으로 기대된다.



## 1.4 연구범위

2009년 상반기 보험 종류별 적발 현황을 살펴보면 적발금액 기준으로는 자동차보험이 67.6%(987억 원), 생명보험의 보장성보험 15.2%(223억 원), 손해보험의 장기보험 10.3%(150억 원) 순으로 나타났다. 적발인원 기준으로는 자동차보험이 87.1%(19,867명), 장기보험 6.7%(1,533명), 보장성보험 5.0%(1,145명)을 차지하고 있다. 이처럼 자동차보험의 비중은 자동차 이용이 생활화되고 다양한 형태의 보험사기가 가능함에 따라 매년 증가하고 있다. 보험사기가 급증하는데 큰 영향을 미치고 있는 소득기반 취약 계층의 생계형 보험사기 중 자동차보험 사기 관련 인원이 89.9%(6,160명)를 차지하고 있다. 자동차보험의 특성상 보험 계약자 외에 피해자 등 다수의 관련자가 존재하므로 보험사기 적발 시 다른 보험사기에 비해 적발 인원이 많은 편이다.

자동차 이용이 일상적인 현대생활에서 자동차 사고는 보험사고 유발이 용이하므로 부당하게 보험금을 편취하는 보험사기에 쉽게 노출되고 있다. 자동차 보험사기 형태로 보면 가피공모 교통사고, 피해자 끼워 넣기, 운전자 바꿔치기(음주운전, 무면허 운전), 사고 발생 후 보험가입 등이 있다.

본 논문에서는 보험사기 비율이 압도적으로 높은 자동차보험 사기를 대상으로 보험사기 적발을 위한 스코어모형을 연구하였다.



## 제 2 장 보험사기 현황

연구에 의하면 세계적으로 지급보험금의 3~15%가 보험사기에 의한 보험금 손실로 추정되고 있으며, 파산한 손해보험사중 30%, 경영위기 생명보험사의 8%가 보험사기로 재무적 위기가 초래된 것으로 나타나고 있다. 본 장에서는 보험사기로 인한 보험금 손실의 증대로 보험사기 적발업무의 체계적인 관리 및 과학적인 대처의 필요성이 매우 커지고 있음을 확인하고 보험사기의 사례를 간단히 살펴본다.

### 2.1 보험사기 현황

#### 2.1.1 해외현황

미국 보험범죄방지국(National Insurance Crime Bureau)에 따르면 최근 미국에서 경기악화로 인해 2009년 상반기 보험사기로 의심되는 손해보험금 지급청구 건수가 지난해에 비해 크게 증가했다고 밝혔다. 미국의 보험사기는 매년 800~2000억 달러로 추산되는데, 가구당 950달러를 추가 부담했다고 한다. 은행이나 모기지 사기 등도 늘고 있기 때문에 한정된 수사관들이 보험사기 이외의 조사에 투입되면서 보험사기에 대한 조사가 느슨해지고 있는데다, 새로운 형태의 보험사기 기법까지 등장하고 있다고 밝혔다.

보험범죄방지국에 따르면 거의 모든 손해보험부문에서 보험사기로 의심되는 보험금지급 청구가 증가했으며, 특히 차량방화와 유리파손의 경우는





각각 20%, 76% 증가했고, 제조물책임보험의 경우에는 90%나 증가했다고 발표했다. 2009년 상반기 보험범죄방지국에 조사의뢰된 보험금지급 청구 건수는 4만 1619건으로 지난해 상반기 3만 6743건에 비해 13% 증가했다. 보험사기가 의심되는 사례가 몇몇 부문에서는 감소했지만 전체적으로는 증가했고 특히 캘리포니아에서는 자동차 방화나 자동차 도난과 관련된 보험사기 의심사례가 많이 나타났다.

선진 외국의 경우 오래전부터 보험사기의 심각성을 인식하고 안정적인 제도 및 수사시스템을 가동해 효율적으로 대처하고 있다. 예를 들면 미국의 경우 1994년에 연방보험사기방지법을 제정하였고, 대부분의 주에서 주 보험사기방지법을 제정하여 보험사기에 대해 강력하게 대처하고 있다.

보험범죄에 대한 수사시스템도 매우 전문화되어 있다. 미국은 주경찰청에 자동차범죄과 등을 두어 자동차보험 사기나 화재보험사기를 전담해 수사하고 있으며, 영국의 경우 경찰청 경제범죄국 산하 사기수사팀에서 보험사기를 전담수사하고 있다(보험범죄방지국, 2009).



### 2.1.2 국내현황

금융감독원이 2009년도 상반기 보험사기 조사실적을 분석한 결과에 따르면, 보험사기 적발 실적은 1,460억 원(22,801명)이며 이는 전년 동기 대비 금액기준 33.6%(367억 원), 인원기준 44.0%(6,969명) 증가하였다.

(단위 : 백만 원, 명, %)

구 분		'06.상반기	'07.상반기	'08.상반기 (a)	'09.상반기 (b)	증감률 (b/a)
적발금액	생보	18,457	19,236	18,430	23,581	28.0
	손보	63,139	81,552	90,844	122,441	34.8
	계	81,596	100,788	109,274	146,022	33.6
적발인원	생보	1,097	1,034	767	1,227	60.0
	손보	10,005	14,016	15,065	21,574	43.2
	계	11,102	15,050	15,832	22,801	44.0

표 2-1 보험사기 적발 현황

보험사기로 적발된 혐의자 연령대별로 살펴보면, 40대가 27.8%(6,334명)로 가장 높고, 30대 25.8%(5,868명), 20대 19.9%(4,542명) 등의 순이다. 10대 및 20대의 증가율은 전년 동기 대비 각각 157.1%, 64.3%로서 다른 연령대(30~60대)의 증가율보다 훨씬 높다. 특히, 상대적으로 범죄의식이 약한 10대의 보험사기는 과거 단순 가담 형태에서 최근에는 학교 선후배 등과 공모하여 조직적이고 계획적으로 보험사기를 실행하는 형태로 발전하여 사회적 문제로 대두되고 있다.

(단위 : 명, %)

구 분	'06년 상반기		'07년 상반기		'08년 상반기		'09년 상반기	
	관련자	구성비	관련자	구성비	관련자	구성비	관련자	구성비
10대	98	0.9	259	1.7	189	1.2	486	2.1
20대	2,124	19.1	2,952	19.6	2,765	17.5	4,542	19.9
30대	3,118	28.1	4,467	29.7	4,396	27.8	5,868	25.8
40대	3,337	30.1	4,305	28.6	4,761	30.1	6,334	27.8
50대	1,845	16.6	2,286	15.2	2,797	17.7	4,221	18.5
60대	485	4.3	653	4.3	788	4.9	1,120	4.9
70대	75	0.7	81	0.5	124	0.7	210	0.9
기타	20	0.2	47	0.4	12	0.1	20	0.1
합계	11,102	100	15,050	100	15,832	100	22,801	100

표 2-2 연령대별 보험사기 적발 인원 현황

보험개발원 보고서에 따르면 지난해 보험사기로 새나간 보험금은 1조 3081억 원이고 고지위반에 따른 누수보험료까지 합하면 1조9000억 원에 달한다고 한다. 이와 관련, 보험연구원은 보상청구 건수 대비 보험사기 건수 비율은 미국, 캐나다, 영국 등과 비슷한 수준이라고 밝혔다. 보험사기로 인해 보험금이 새나갈 경우, 그 부담은 선량한 보험가입자들이 떠안게 된다.

우리나라의 보험사기 규모는 2조 2303억 원으로 추정되고, 누수 보험금 및 보험료를 4인 가구 기준으로 계산하면 보험사기로 인해 가구당 보험료가 매달 14만원이 늘어나는 셈이다(보험개발원, 2009).

생명보험에서 사기로 추정되는 금액 중 5.7%만이 적발되고 있으며, 손해 보험 중에는 16.2%만이 적발되고 있는 실정이다. 연간 손해보험 사기추정 금액 중 83.8%(9,702억 원), 생명보험 사기추정금액 중 94.3%(10,111억 원)의 보험금이 부당하게 지급되어 보험회사 및 선의의 보험자들에게 피해를 주고 있다.



(단위 : 억 원, %)

구 분	사기추정금액	적발금액	적발금액(%)
손해보험	11,580	1,878	16.2 %
생명보험	10,723	612	5.7 %
합 계	22,303	2,490	11.2 %

표 2-3 보험사기 추정 금액 현황

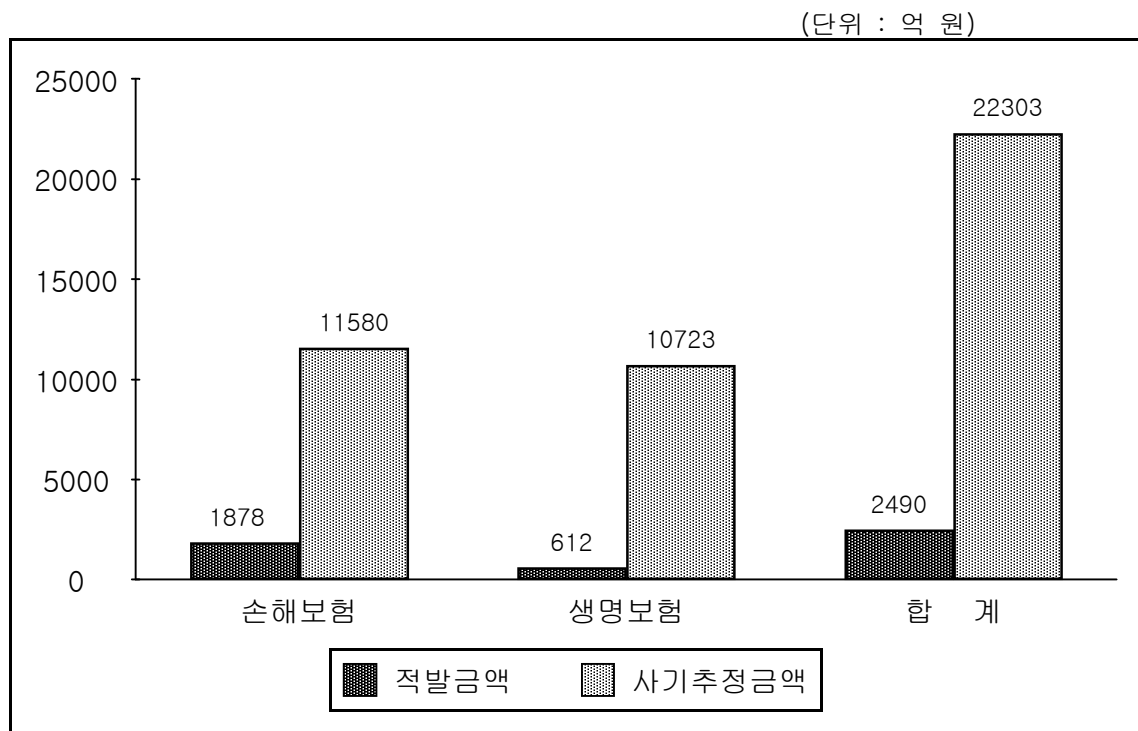


그림 2-1 보험사기 추정 금액 현황

손해보험의 2008년 보험사기 적발 실적을 살펴보면, 자동차보험의 사기 적발 비율이 월등하게 높다. 적발금액 기준으로 약 80%이상, 적발인원 기준으

로 약 90%이상이 자동차 손해보험에서 발생하는 보험사기이다. 이런 현상은 손해보험은 보험계약자 외 피해자 등 다수의 관련자가 존재하는데 기인한다.

(단위 : 백만 원, %)

구 분	'06년		'07년		'08년		'09년	
	적발금액	구성비	적발금액	구성비	적발금액	구성비	적발금액	구성비
자동차보험	104,788	95.7	123,897	89.2	135,882	82.3	177,912	81.6
장기보험	2,562	2.3	12,765	9.2	23,263	14.1	32,180	14.8
화재보험	1,221	1.1	280	0.2	2,280	1.4	4,297	2.0
상해보험	220	0.2	135	0.1	2,387	1.4	2,995	1.4
특종보험	704	0.6	1,788	1.3	1,379	0.8	496	0.2
기 타	-	-	-	-	-	-	40	0.0
계	109,495	100	138,865	100	165,191	100	217,920	100

표 2-4 손해보험 종류별 보험사기 적발 금액 현황

(단위 : 명, %)

구 분	'06년		'07년		'08년		'09년	
	적발인원	구성비	적발인원	구성비	적발인원	구성비	적발인원	구성비
자동차보험	17,649	97.3	22,565	93.3	24,825	86.8	35,852	91.2
장기보험	349	1.9	1,440	6.0	3,387	11.8	3,119	7.9
화재보험	13	0.1	6	0.0	18	0.1	21	0.1
상해보험	9	0.0	22	0.1	238	0.8	285	0.7
특종보험	119	0.7	151	0.6	143	0.5	43	0.1
기 타	-	-	-	-	-	-	2	0.0
계	18,139	100	24,184	100	28,611	100	39,322	100

표 2-5 손해보험 종류별 보험사기 적발 인원 현황



## 2.2 보험사기 유형

보험사기란 보험가입자가 보험자로부터 위법적인 방법으로 거액의 보험금을 편취하는 모든 행위를 말하는 것으로 크게 경성사기와 연성사기 두 가지로 분류할 수 있다.

### (1) 경성사기(Hard Fraud)

경성사기란 고의적이고 악의적인 범죄를 말하는 것으로 범죄 수법이 매우 치밀하고 사전 계획적인 사기를 말한다. 보험 증권에서 담보하는 재해, 상해, 도난, 방화 등의 손실을 의도적으로 각색하거나 조작하는 행위가 해당된다. 예를 들면, 계획적으로 피해자와 피의자가 공모하여 고의로 자동차 사고를 유발하여 보험금을 부당 수취하는 경우이다.

### (2) 연성사기(Soft Fraud)

연성사기란 의도적이거나 사전 계획된 것이 아니라 정상적인 사고 발생 후 개인적인 이익을 위해 우발적으로 사고를 과장되게 부풀리는 행위의 사기를 말하는 것으로 기회주의적 사기라고도 한다. 예를 들면, 자동차 사고를 당했을 때, 기존의 병력을 숨기거나 사고와 무관한 부분까지도 포함시켜서 거액의 보험금을 청구하는 경우이다.



## 2.3 보험사기 사례

### (1) 고교생 등으로 구성된 교통사고 위장 보험사기

서울 OO구 소재 고등학교 선후배인 A(22세, 남)와 OO구 소재 고등학교 선후배인 A(22세, 남)와 B(18세, 남) 등 4명은 교통사고 보험사기를 목적으로 함께 범행할 고등학교 선후배 및 동창생 104명을 모집하여, 이들에게 ‘첫째. 사고가 나면 무조건 입원을 하라.’, ‘둘째. 같은 병원에 두 번 이상 입원을 하지 말라.’, ‘셋째. 경찰에 걸리면 무조건 사고라고 우겨라.’라는 행동지침을 교육시킨 후, ‘07년 1월부터 ’08년 3월까지 일방통행로·편도 1차로 등에서 고의 또는 허위로 교통사고를 내고, 보험회사로부터 입원 치료비 등 명목의 보험금을 타내는 수법으로 총 100여회에 걸쳐 4억 원 가량을 편취하다 수사기관에 의해 적발되었다.

(서울지방경찰청 외사과 ‘09.6.29 발표)

### (2) 여중생을 끌어들인 고의 교통사고

A(19세, 남) 등 7명은 지역 고등학교 선·후배 사이로, 오토바이 폭주족으로 단속되어 부과 받은 벌금(개인별 200~300만원)을 납부하기 위해서 B(16세, 여) 등 여중생들을 끌어들여 다단계 형 보험사기단을 결성(53명)하고, ‘07년 9월 승용차와 오토바이를 나눠 타고 마포구 소재 노상에서 자신들의 차량과 오토바이를 고의 추돌 후, 상해진단서를 발부받아 보험사로부터 합의금 명목으로 4백만 원을 편취하는 등 ‘06년 10월부터 ’08년 12월



까지 서울·경기 일대에서 위와 같은 수법으로 총 70회에 걸쳐 3억 원 상당의 보험금을 편취하다 수사기관에 의해 적발되었다.

(서울지방경찰청 광역수사대 '09.7.28 발표)

### (3) 기왕증을 상해사고로 가장한 보험사기

서울 강서구에 거주하는 송○○(64년생)은 2007년 1월 초순경 논산-천안 간 고속도로에서 발생한 교통사고로 서울 ○○병원에서 '요추4번~천추1번 간 추간판탈출증' 이라는 진단으로 치료를 받고 동 병원으로부터 2007년 8월경에 “척추장해3급9항의 장해 진단서”를 발급 받아 K·X 보험사에 228백만 원 상당의 장해보험금을 청구하여 73백만 원(X 보험사)을 편취하고 155백만 원(K 보험사)은 미수에 그쳤다. 경찰청과 공조 조사를 진행한 결과, 다음과 같은 범죄증거를 확보하여 2008년 5월경에 혐의자 송○○를 불구속 기소되었다.

#### <주요 보험사기 혐의 증거>

- 혐의자가 판정받은 위 장해등급 제3급9항은 무거운 물체를 전혀 들지 못하는 정도의 허리장해임에도 혐의자가 무거운 이삿짐을 옮기는 장면이 목격된 점
- 국내 유명 종합병원에서 위 장해진단에 대한 의료자문을 받은 결과 혐의자의 병증이 “퇴행성 질환”이라는 소견을 보인 점
- 혐의자가 위 장해진단을 받기 수일 전에 나이트클럽에서 발생한 폭행사고에 직접 가담하여 벌금형을 선고 받은 점





## 제 3 장 데이터마이닝 기법의 적용

점차 지능화되는 보험사기를 적발하기 위해서는 보험사기 적발 업무에 과학적 접근이 필요하다. 본 장에서는 다양한 데이터마이닝 기법들로 보험사기 적발 모형을 생성할 때 적용상의 특징을 간단히 살펴본다.

### 3.1 데이터마이닝 기법

#### 3.1.1 Decision Tree

Decision Tree는 의사결정규칙을 나무구조로 도식화하여 분류나 예측을 수행하는 분석방법이다. Decision Tree 모형 생성 시 특징은 다음과 같다.

- (1) 변수선택이 가능하여 모형 생성이 용이하다.
- (2) 모형의 결과를 규칙형태의 도식화가 가능하여 모형에 대한 전문적인 지식이 없더라도 다른 모형 생성 기법들에 비해 해석이나 이해가 쉽다.
- (3) 복잡한 입력변수와 목표변수간의 관계 모형화가 가능하고 결측값을 하나의 범주로 간주하여 모형화가 가능하다.
- (4) 나무 구조의 depth가 너무 깊은 경우에는 예측력의 저하뿐 만 아니라 해석도 어려워 질 수 있다.

Decision Tree 기법은 모형의 정확도 보다는 분석과정이나 결과에 대한 설명이 필요한 경우에 더욱 유용하게 사용될 수 있다.



### 3.1.2 Logistic Regression

Logistic Regression은 두 개의 값을 가진 목표변수와 설명변수들 사이의 인과관계를 분석하는 기법이다. 목표변수가 이원적인 경우를 흔히 비선형 회귀분석이라고 한다. 각 변수의 구간대별 스코어카드 생성이 가능한 기법으로 보험사기 적발 모형 생성 시 일반적으로 이용되는 방법으로 그 특징은 다음과 같다.

- (1) 각 변수의 구간대별 스코어카드 생성이 가능하다.
- (2) 회귀계수와 오즈비를 이용하여 입력변수와 목표변수와 관계를 쉽게 알 수 있어 해석이 편리하다.
- (3) Stepwise Selection, Forward Selection, Backward Elimination 등의 방법으로 불필요한 변수를 제거하고 필요한 소수의 변수만을 속아낼 수 있고, 이를 통하여 모형의 예측력과 해석력을 높일 수 있다.

본 논문에서는 Fraud/Non-Fraud(보험금 부당 청구건/보험금 정상 청구건)라는 두 개의 값을 가지는 목표변수와 독립 변수들 간의 인과관계를 Logistic 함수를 이용하여 추정하였다.

### 3.1.3 Neural Network

Neural Network는 인간 두뇌의 신경망 구조를 모방하여 만들어진 기법으로 여러 개의 뉴런들이 상호 연결하여 반복적인 학습 과정을 거쳐 데이터에 숨어있는 패턴을 찾아내고 입력에 상응하는 최적의 출력 값을 예측한



다. Neural Network 기법으로 모형 생성 시 특징은 다음과 같다.

- (1) Neural Network 모형은 Logistic Regression이나 Decision Tree 등 다른 예측 기법들에 비해 매우 뛰어난 예측력을 가진다.
- (2) 모형 생성 시 소요되는 시간이 많으며 과잉적합의 위험이 존재한다.
- (3) 입력변수와 목표변수와의 관계를 파악하는 것이 거의 불가능하며 모형의 결과에 대한 해석이 어렵다.
- (4) Neural Network 기법은 그 모형의 복잡성으로 인하여 입력변수의 선택에 따라 매우 민감하게 반응하므로 성공적인 모형 생성을 위해서는 변수 선택 과정이 매우 중요시 된다.

Neural Network는 매우 복잡한 구조를 가진 방대한 데이터 사이의 연관 관계나 패턴을 찾아내고 그를 통하여 적중률 높은 예측을 원하는 경우에 유용하다. 어떤 분야에서는 결과의 규칙을 설명할 수 있느냐가 모형의 실무 적용 여부 판단에 기준이 되기도 한다. 따라서 Neural Network 모형은 결과로 스코어가 생성되는 과정을 설명하는 것 보다 적중률 높은 결과값 그 자체가 중요시될 때 선호된다.

자동차 상해보험 모형 구축 시 Neural Network 모형이 Decision Tree나 Logistic Regression등의 모형보다 우수한 성능을 보인다는 것은 다른 연구에서 이미 증명된바 있다(Stijn Viaene et al., 2002).



### 3.1.4 Ensemble

앙상블은 다수의 단일 모델을 만들고 각기 다른 단일 모델들을 여러 가지 방법으로 결합하여 하나의 새로운 통합 모델을 구성하는 방법을 말한다.

앙상블 학습을 하는 이유는 다양한 단일 모델들의 결과를 결합함으로써 좀 더 안정적이고 신뢰성이 높은 결과를 얻을 수 있고 예측 성능이 우수하기 때문이다. 또한 단일모델 하나가 모든 패턴에 대해서 좋은 성능을 발휘할 수 없기 때문에 다양한 모델들을 사용하게 되는데 앙상블 모델이 단일 모델 하나만 사용했을 때 보다 성능 향상을 보인다는 것은 이미 다른 연구에서 여러 차례 증명되었다.

본 논문에서는 다른 선행 연구들과 같이 보험사기 사례에서도 앙상블 모델이 단일모델보다 안정적이고 우수한 예측력을 가짐을 실험을 통해 증명하였다. 앙상블 모델은 이해가 쉽고 구현이 간단한 Simple Average를 이용하여 생성하였다.

$$f_{ensemble} = \frac{1}{n} \sum_{i=1}^n f_i(x) \quad \text{where. } i=1 \dots n (n: \text{단일 모델의 개수})$$

### 3.1.5 Link Analysis

Byrne(2002)에 의하면 ‘Link Analysis는 개체간의 관계에서 발생한 데이터 속에서 패턴 또는 움직임 추출하기 위해 서로 연관되어 되어 있는 개체들의 관계를 네트워크화 하는 과정’이라 정의하였다. Link Analysis가



다양한 분야에 사용되는 것은 대용량 데이터 상에서 각 Entities간의 연결 관계를 탐색하고, 명확히 정의하는 수단으로 매우 효과적이기 때문이다.

대부분의 과학적 분석기법은 정량적 기법 즉, 데이터중심적인 접근방법이 주를 이루고 있었지만 이러한 방법들은 문제의 영역(domain)이 비체계적이고 문제나 대상이 수시로 변형, 발전되는 경우 특정 패턴을 파악하기가 매우 어렵다(Davison, 1993). 보험사기 적발에서도 마찬가지로 단순한 보험사기가 아닌 계획적인 공모에 의한 보험사기의 경우 변수의 가중치로 추정하게 되는 Logistic Regression이나 Neural Network로는 추적하기가 어렵다.

Link Analysis는 보험사기 혐의자 적발에 매우 효과적이다. 그러나 SIU(special investigation unit) 기능을 대체할 수는 없다. 링크분석은 축적된 관계데이터를 통해서 보험사기 혐의자를 수 십분 이내에 추출하는 뛰어난 속도와 상당한 정확성을 가지고 있지만, 보험사기 혐의자를 보험사기자라고 단정할 만한 구체적인 법적 증거를 확보하는 업무와는 별개다. 즉, 보험사기 혐의자를 집중 조사해서 궁극적으로 보험사기자로 결정하는 것은 SIU의 고유 업무이다. 따라서 링크분석이 SIU와 유기적으로 연계될 때 그 효과가 극대화될 수 있을 것이다(김현수, 2003).



### 3.2 스코어카드 생성

스코어 모형에서 얻은 결과 값을 이용해 스코어카드를 작성하면 모형에 사용된 변수들 각각의 영향력의 정도를 파악하는데 유용하다. 본 논문에서는 Logistic Regression의 수행 결과로 얻어진 추정치들을 이용하여 다음과 같은 순서로 스코어카드를 작성해보았다.

스코어카드는 Logistic Regression을 이용하여 산출된 스코어를 각 변수 범주에 할당하여 생성된다.

#### (1) 스코어 Scale 조정 범위 설정

스코어의 해석을 용이하게 하기 위하여 스코어의 Scale을 조정하기 위해 먼저 생성될 스코어의 최소값과 최대값을 정해준다.

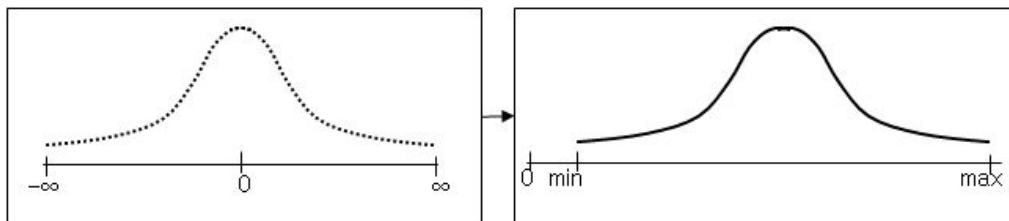
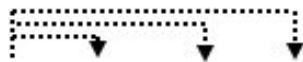


그림 3-1 스코어 Scale 조정 범위 설정

## (2) 절편(Intercept) 배분

절편(Intercept)의 영향이 각 변수들에 동등하게 영향을 미치도록 절편 값을 설명변수의 개수로 나누어 각 변수 범주별 스코어 값에 배분해 준다.


$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

## (3) 스코어 양수화

해당 변수의 범주별 스코어들 중 최소값을 해당 변수의 범주별 스코어 값들에서 빼주어 음의 값을 가지는 스코어들을 양의 값으로 바꾸어 주어 스코어 해석을 쉽게 해준다.

## (4) 스코어 생성

조정하고자 했던 스코어의 범위로 Scale을 조정하고 각 변수 범주별로 스코어를 할당해 스코어 카드를 생성한다. 각 변수들의 범주별 스코어를 산출하는 Scale 조정 식은 다음과 같다.

$$\text{Scaled Score} = \text{양수화된 Score} \times \frac{1}{M} \times (\max - \min) + \min \times \frac{1}{p}$$

Where. M : 스코어카드에서 얻어지는 스코어들의 변수별 최대값

max : 조정된 후 산출되는 스코어의 최대값

min : 조정된 후 산출되는 스코어의 최소값

p : 입력변수의 개수



		①	②	③	④	⑤	⑥	⑦		
Parameter	Estimate	절편의 배분	변수별	변수	변수별	스코어 최대화	최소값배분	Score	변수별	
Intercept	-69.7696	-4.3606	최소값	양수화	최대값	0.3	6.25		Score max	
GRP_Variable1	1	30.9162	26.5556	-4.3606	30.9162	30.916			16	16
GRP_Variable1	2	7.8714	3.5108		7.8714				9	
GRP_Variable1	3	0.4313	-3.9293		0.4313				6	
GRP_Variable1	4	11.0924	6.7318		11.0924				10	
GRP_Variable1	5	0	-4.3606		0				6	
...										
GRP_Variable10	1	-20.2543	-24.6149	-86.3066	61.6917	81.946			25	31
GRP_Variable10	2	-10.27	-14.6306		71.676				28	
GRP_Variable10	3	-40.2226	-44.5832		41.7234				19	
GRP_Variable10	4	-81.946	-86.3066		0				6	
GRP_Variable10	5	0	-4.3606		81.946				31	

변수별 Score Max 합 : 약 1,000점

그림 3-2 스코어카드 생성 예

① : 절편의 배분

$$\text{변수당 범주별 } Estimate + \frac{Intercept}{\text{총 입력변수 개수}}$$

② : 변수 별 최소값 = Min(변수당 범주별 ①의 값)

③ : 변수 양수화 = ① - ②

④ : 변수 별 최대값 = Max(변수당 범주별 ③의 값)

⑤ : 스코어 최대화 = 950 / Sum( Max(변수당 범주별 ③의 값))

⑥ : 스코어 최소값 배분 = 50/총 입력변수 개수

⑦ : 변수당 범주 별 스코어 = ROUND(③\*⑤+⑥)





## 제 4 장 앙상블 효과 및 스코어카드 생성

### 4.1 실험주제

보험사기 적발 모형 생성 시 각 변수의 범주별 스코어카드 생성이 가능한 Logistic Regression 방법이나 모델의 결과를 규칙형태의 도식화가 가능한 Decision Tree 기법을 사용하여 모형을 생성한 후 효과성을 비교 토 하여 최종 모델을 결정하는 것이 일반적인 방법이다.

단일 모형 하나를 이용하는 것 보다는 앙상블 모형을 구성하였을 때 성능이 훨씬 안정적이고 우수한 것은 알고 있지만, 모형의 스코어에 대한 설명이 중요시 되는 보험사기 적발 업무에서는 해석의 어려움 때문에 앙상블 모형이 실제업무에 이용되지는 않고 있다.

본 장에서는 다양한 데이터마이닝 기법들을 이용하여 보험사기 적발을 위한 여러 개 단일 모형을 생성한 후, Simple Average 방법으로 앙상블 모형을 구성하여 성능을 비교하여 우수한 앙상블 모형의 성능을 확인한다. 효과가 입증된 앙상블 모형의 스코어가 설명력이 중요시되는 보험사기 적발 업무에서 효율적으로 사용될 수 있도록 스코어카드를 생성하여 모형의 스코어 해석에 도움이 될 수 있는 방법을 제시한다.



## 4.2 실험개요

### 4.2.1 데이터이해

본 논문에서는 자동차 보험 부문 모형을 생성해 보았다. 자동차 보험사가 적발 업무에 모형이 적용될 수 있는 단계는 접수, 계약 및 사고조사, 피사고, 치료관리 총 4단계로 나누어 볼 수 있다. 각 단계마다 모형 적용이 가능하나 본 연구에서는, 모형 수립에 사용 가능한 정보의 양이 충분한 피해조사 단계에서 적용 가능한 모형을 생성하여 보았다. 생성되는 모형은 보험료 청구 대상 별로 대인/대물/자손/자차 총 4가지 종류로 구분하여 모델링을 실시하였다. 피해 대상이 사람인 경우 상대방 차량의 피해자는 대인, 자신 차량의 피해자는 자손 담보로 정의되며, 피해 대상이 사물인 경우 상대방의 피해물은 대물, 자신의 피해물은 자차로 구분된다.

#### (1) 데이터 정의

실험에 이용된 데이터는 2007년 2월부터 2008년 3월까지 발생한 국내 A 보험사의 일반적인 자동차 보험 청구 건을 수집하여 사용하였다. 수집된 데이터를 Train용과 Test용으로 구분하여 사용하였다.

Train용으로 2007년 2월부터 2008년 2월까지 총 12개월간 발생한 보험료 청구 건을 사용하였고, Test용으로 2008년 3월 한 달간 발생한 보험료 청구 건을 사용하였다.



## (2) 목표변수 정의

자동차 보험료 청구 건 중 일어나지 않은 사고를 허위로 청구하는 보험 사고를 가공한 경정사기, 발생한 사고와는 상관없는 사고 피해를 과장하는 연성사기 등 보험료를 부당으로 청구 하는 모든 건이 목표변수가 된다.

## (3) 입력 변수 정의

보험사기 적발 모형에서 사용된 입력변수는 Yes/No 또는 1/0으로 표현 되는 보험사기 지표변수(Reg Flag Indicator)와 일반 Predict 변수가 있다.

### - 보험가기 지표변수(Red Flag Indicators)

보험사기 관련자의 신상, 사고 상황, 보험종목 등에 대해 축적된 통계 자료를 이용하여 만들어진다. 보험사기 관련 정보들을 정형화하고 표준화하여 공통적으로 발견되는 사항을 지표화한 것으로 보험사기 가능성에 대한 실마리를 제공한다. 이러한 지표변수들에 해당되는 변수로는 최초 경찰 신고 여부(Y/N), 기명피보험자와 자차 운전자 불일치 여부(Y/N) 등 이 사용된다.

### - 일반 Predict 변수

피해자의 연령이나 동승인원수, 동승자 연령대, 최초 사고 발생 시간대등 보험료가 청구된 자동차 사고와 관련하여 얻을 수 있는 기본적인 정보들이 해당된다.



자동차 상해보험 모형 구축 시 Indicator 변수만 있을 때 보다, Predictor 변수를 추가하여 하였을 때 모형의 성능이 향상된다는 것은 다른 연구에서 이미 증명된바 있다(Stijn Viaene et al., 2002). 본 논문에서는 데이터준비 단계에서 제시된 방법으로 선택된 Indicator와 Predictor 변수만이 최종 모형의 입력변수로 사용되었다.

#### 4.2.2 데이터준비

##### (1) 데이터 샘플링

분석의 용이성을 높이기 위하여 보험 청구건 전체 중 일부만을 샘플링 하여 분석을 수행한다. 샘플링 데이터는 전체 모집단의 특성을 반영할 수 있도록 대표성을 가지고 있어야 하며 원활한 분석 수행이 가능하도록 분석 환경을 고려한 적절한 수준의 데이터 사이즈로 샘플링 되어야 한다. 또한 발생한 결과 해석 시 계절성 등의 이벤트적 영향이 배제되고 일반적 의미 부여가 가능하도록 샘플링이 수행되어야 한다.

본 논문에서는 Fraud건은 모집단 내의 전수 데이터를 사용하고 Non-fraud건 만 샘플링 하는 Over-Sampling 방법으로 상대적으로 그 건수가 적은 Fraud 정보의 손실을 줄일 수 있도록 하였다. 모집단의 분포를 유지하는 일반 샘플링을 수행하면 상대적으로 발생 비율이 낮은 Fraud의 특성은 나타나지 않거나 샘플링에서 제외되는 경우가 발생할 가능성이 있기 때문이다. 샘플링 된 데이터 내에 Non-Fraud와 Fraud의 비율은 경험적으로 볼 때 모델 학습에 가장 효과적인 9:1의 비율로 샘플 데이터를 구성하였다. 그 비율은 대인/대물/자손/자차 각각 모형에 동일하게 적용하였다.



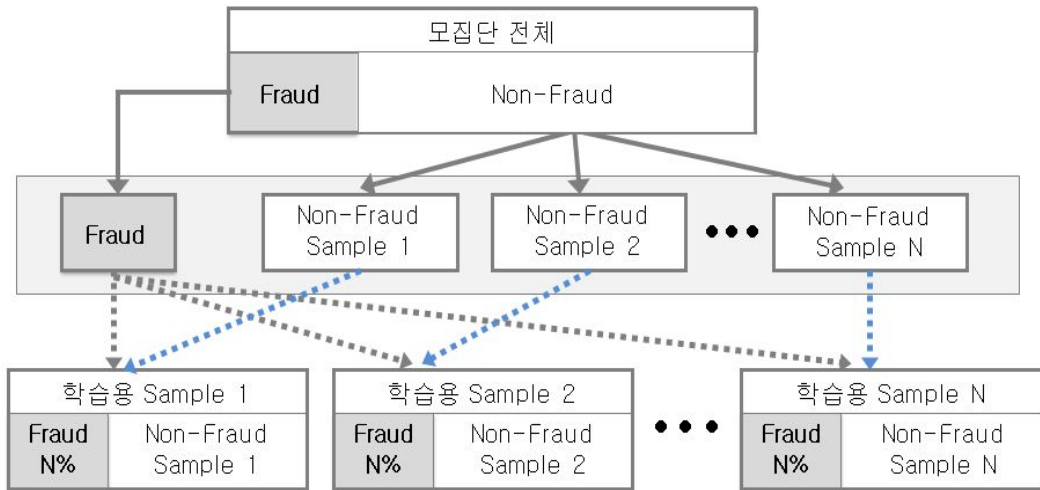


그림 4-1 학습용 샘플 데이터의 구성

구분	Train			샘플링		Test		
	청구건	Fraud	Fraud%	샘플건	정상:사고	청구건	Fraud	정상:사고
대인 모델	45,023	1,209	2.69%	10,081	9 : 1	15,176	71	214 : 1
대물 모델	66,914	1,015	1.52%	9,094	9 : 1	28,054	79	355 : 1
자손 모델	37,180	531	1.43%	5,247	9 : 1	15,679	42	373 : 1
자차 모델	7,853	657	8.37%	4,510	9 : 1	954	43	22 : 1

표 4-1 학습용 샘플 데이터 생성

샘플 데이터의 대표성을 검증하기 위하여 PSI(Population Stability Index) 지표를 이용하였다. PSI 지표는 다음 수식을 이용하여 구할 수 있다.

$$\sum \left\{ (NonFraud\% - Fraud\%) \times \ln \left( \frac{NonFraud\%}{Fraud\%} \right) \times 100 \right\}$$

사고월, 지급준비금, 피보험자의 연령대 변수의 분포를 비교하여 각 모형의 대표성을 검증하였다. 각 모형에서 두 집단 간의 차이를 나타내는 PSI 지표가 3개 변수에서 모두 10 미만으로 나타나므로 두 집단의 분포가 동일함을 의미하며 샘플 데이터가 전체 모집단의 분포를 잘 반영하고 있다.

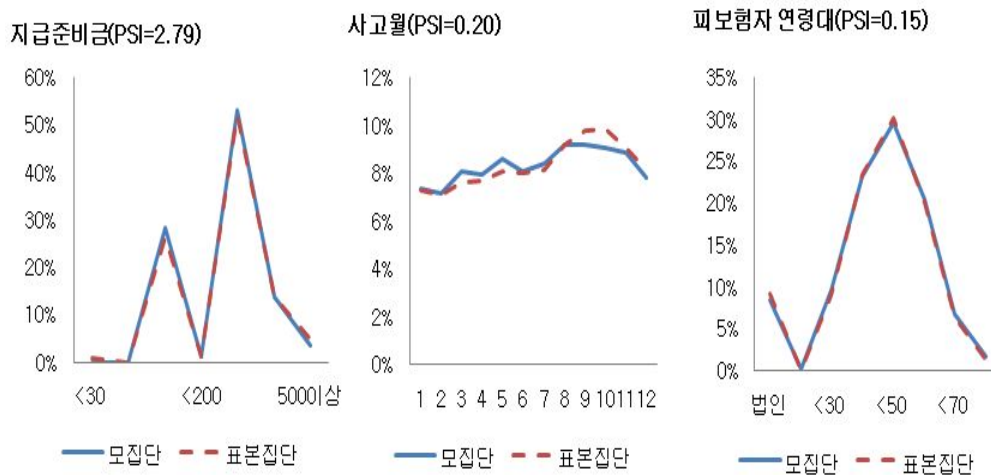


그림 4-2 대인모형 샘플 데이터 대표성 검증

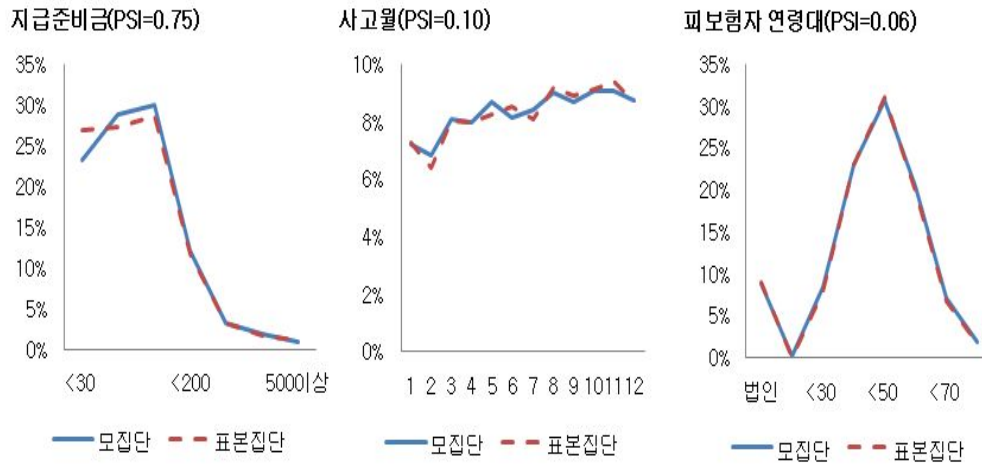


그림 4-3 대물모형 샘플 데이터 대표성 검증

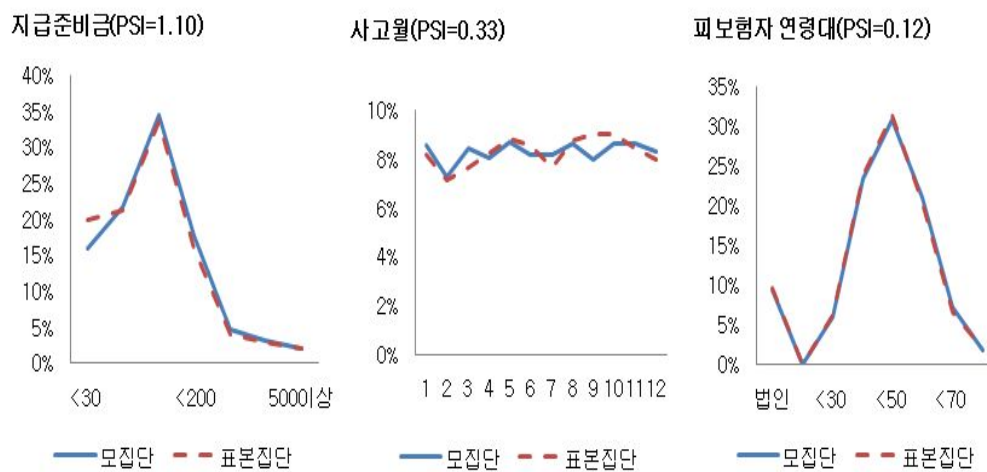


그림 4-4 자차모형 샘플 데이터 대표성 검증

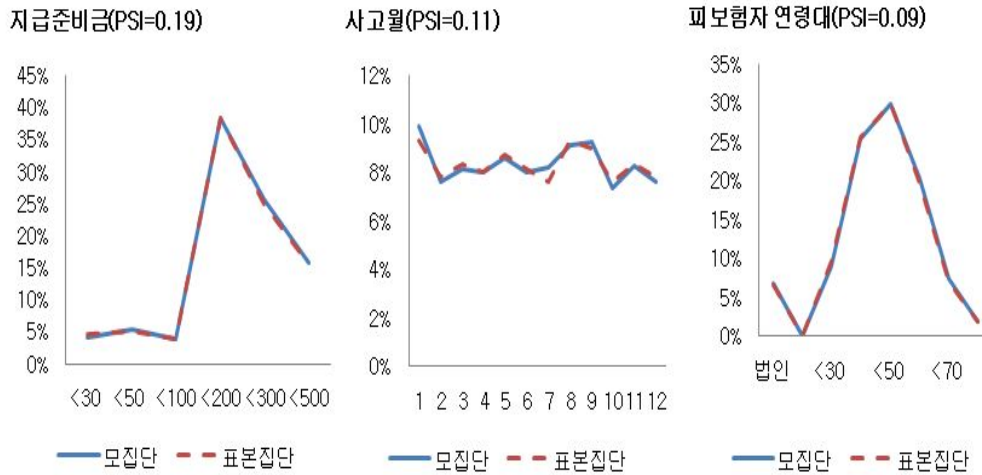


그림 4-5 자손모형 샘플 데이터 대표성 검증

## (2) 입력 변수 선택

모형에 사용된 최종 입력변수는 IV(Information Value)를 통한 유의성 검증, 재 그룹핑, 상관분석 3단계로 수행하였다. 먼저 생성된 변수 중 각 모형에 유의한 변수를 WOE와 IV를 이용하여 1차적으로 추출하였다.

- WOE(Weight of Evidence) : 목표변수에 미치는 영향도를 측고 있다.

$$WOE = \ln\left(\frac{Non\ Fraud\%}{Fraud\%}\right) \times 100$$

- IV(Information Value) : 해당 변수가 목표변수를 설명하는 정보량을 나타내는 지수이다.

$$IV = \sum (Non\ Fraud\% - Fraud\%) \times WOE$$



일반적으로 IV 값의 범위에 따른 의미는 다음과 같다.

- 10 미만 : Weak Variable
- 10 ~ 30 : Medium Variable
- 30 ~ 50 : Strong Variable
- 50 이상 : Very Strong Variable

1차적으로 추출된 변수의 안정성 및 효율성을 높이기 위하여 그룹이 너무 많거나 변수의 의미를 퇴색시키는 변수 등을 재 그룹을 통하여 최종 입력 변수로 사용될 형태로 변환하였다. 유의성이 높은 변수라도 비슷한 형태의 변수가 투입되면 모델의 안정성 및 가독성을 저하시키므로 상관관계가 높은 변수를 제거하였다.

#### 4.2.3 모형생성

보험사기 적발을 위한 여러 가지 단일 모형을 생성한 후 앙상블 모형을 구성한다. 대인/대물/자차/자손 각각의 경우에 최종 앙상블 모형은 Neural Network 3개와 Logistic Regression 모형 2개로 구성하였으며 Simple Average 방법을 적용하였다.

앙상블 모형이 생성되고 나면 앙상블 모형의 스코어를 10개 구간으로 나누어 상위 10% 구간에 해당되는 스코어를 받은 건들에 Target=1, 나머지 90%에는 Target=0으로 부여하여 새로운 목표변수를 생성한다. 10% 구간에 Target=1을 부여한 것은 단일 모형 학습에 적용된 정상대사고 비율인 9:1을 그대로 적용하기 위한 것이다.



앙상블 모형의 스코어카드를 생성하기 위하여 앙상블 모형에 사용된 모든 입력변수들을 Input으로 하고, 새롭게 생성된 목표변수를 Target으로 Logistic Regression으로 재학습시켜 스코어카드 생성에 필요한 정보인 Intercept와 변수의 범주별 추정치를 구할 수 있다. 앙상블 모형의 스코어 카드는 새로운 목표변수를 이용하여 Logistic Regression을 수행한 후 산출된 스코어를 각 변수의 범주별로 할당하여 3장에서 제시된 방법으로 생성된다.

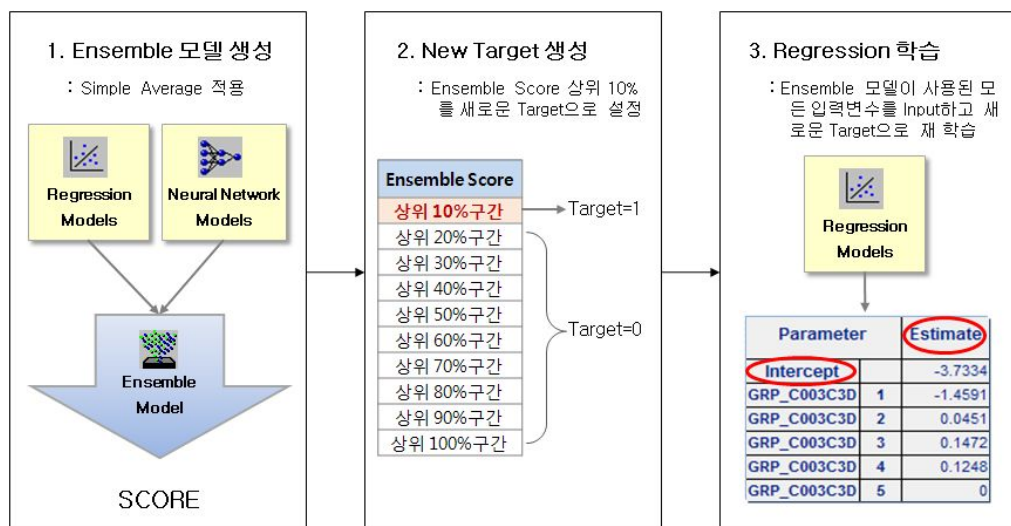


그림 4-6 앙상블 모형의 스코어카드 생성

## 4.3 실험결과

### 4.3.1 성능평가

단일 모형과 앙상블 모형의 성능을 비교한다. 각 모형들이 산출한 스코어를 십분위 분석으로 10개의 구간으로 나눈다. 성능 평가는 스코어 상위 10%를 기준으로 Fraud 검출력과 검출효율을 비교한다. 성능 평가에 이용되는 검출력은 누적 %Response로 표현되며 각 등급에서 목표범주 1의 누적 비율을 나타내고, 검출효율은 누적 %Captured Response로 목표범주 1에 속하는 건들이 각 등급에 얼마나 분포하고 있는지를 나타낸다.

실험결과 대인/대물/자차/자손 각 담보별로 생성된 모형의 성능은 검출력과 검출효율 측면에서 단일모형보다 앙상블 모형에서 더 우수하였다.



스코어 구간	Logistic Regression 단일 모형					앙상블 모형
	모형1	모형2	모형3	모형4	모형5	NN3개+Reg2개
상위 10%	2.36	2.36	<b>2.43</b>	2.36	2.36	<b>2.56</b>
상위 20%	1.67	1.64	1.74	1.67	1.74	1.70
상위 30%	1.27	1.33	1.31	1.33	1.31	1.36
상위 40%	1.07	1.07	1.05	1.07	1.03	1.05
상위 50%	0.89	0.88	0.85	0.88	0.85	0.85
상위 60%	0.74	0.74	0.74	0.75	0.75	0.74
상위 70%	0.65	0.66	0.66	0.65	0.65	0.64
상위 80%	0.58	0.58	0.58	0.58	0.58	0.58
상위 90%	0.52	0.52	0.52	0.52	0.52	0.52
상위 100%	0.47	0.47	0.47	0.47	0.47	0.47

표 4-2 대인 모형 스코어의 %Response 비교

스코어 구간	Logistic Regression 단일 모형					앙상블 모형
	모형1	모형2	모형3	모형4	모형5	NN3개+Reg2개
상위 10%	50.7	50.7	<b>52.1</b>	50.7	50.7	<b>54.9</b>
상위 20%	71.8	70.4	74.6	71.8	74.6	73.2
상위 30%	81.7	85.9	84.5	85.9	84.5	87.3
상위 40%	91.5	91.5	90.1	91.5	88.7	90.1
상위 50%	95.8	94.4	91.5	94.4	91.5	91.5
상위 60%	95.8	95.8	95.8	97.2	97.2	95.8
상위 70%	97.2	98.6	98.6	97.2	97.2	95.8
상위 80%	100.0	100.0	100.0	100.0	100.0	100.0
상위 90%	100.0	100.0	100.0	100.0	100.0	100.0
상위 100%	100.0	100.0	100.0	100.0	100.0	100.0

표 4-3 대인 모형 스코어의 %Captured Response 비교

스코어 구간	Logistic Regression 단일 모형					앙상블 모형
	모형1	모형2	모형3	모형4	모형5	NN3개+Reg2개
상위 10%	0.81	0.82	<b>0.81</b>	0.81	0.79	<b>0.84</b>
상위 20%	0.81	0.82	0.81	0.81	0.82	0.83
상위 30%	0.60	0.61	0.60	0.59	0.62	0.64
상위 40%	0.56	0.58	0.57	0.56	0.56	0.58
상위 50%	0.48	0.48	0.48	0.48	0.48	0.48
상위 60%	0.28	0.28	0.28	0.28	0.28	0.28
상위 70%	0.28	0.28	0.28	0.28	0.28	0.28
상위 80%	0.28	0.28	0.28	0.28	0.28	0.28
상위 90%	0.28	0.28	0.28	0.28	0.28	0.28
상위 100%	0.28	0.28	0.28	0.28	0.28	0.28

표 4-4 대물 모형 스코어의 %Response 비교

스코어 구간	Logistic Regression 단일 모형					앙상블 모형
	모형1	모형2	모형3	모형4	모형5	NN3개+Reg2개
상위 10%	55.7	55.7	<b>55.7</b>	58.2	51.9	<b>58.2</b>
상위 20%	59.5	59.5	59.5	58.2	59.5	60.8
상위 30%	74.7	72.2	73.4	73.4	74.7	77.2
상위 40%	82.3	83.5	83.5	82.3	82.3	83.5
상위 50%	87.3	87.3	87.3	87.3	87.3	87.3
상위 60%	98.7	98.7	98.7	100.0	98.7	100.0
상위 70%	100.0	98.7	98.7	100.0	98.7	100.0
상위 80%	100.0	98.7	98.7	100.0	98.7	100.0
상위 90%	100.0	98.7	98.7	100.0	98.7	100.0
상위 100%	100.0	100.0	100.0	100.0	100.0	100.0

표 4-5 대물 모형 스코어의 %Captured Response 비교

스코어 구간	Logistic Regression 단일 모형					앙상블 모형
	모형1	모형2	모형3	모형4	모형5	NN3개+Reg2개
상위 10%	0.79	0.82	<b>0.85</b>	0.85	0.79	<b>0.86</b>
상위 20%	0.66	0.66	0.67	0.63	0.66	0.74
상위 30%	0.64	0.64	0.61	0.64	0.63	0.63
상위 40%	0.49	0.49	0.47	0.51	0.49	0.49
상위 50%	0.46	0.47	0.47	0.46	0.46	0.45
상위 60%	0.41	0.40	0.40	0.39	0.39	0.41
상위 70%	0.36	0.36	0.36	0.37	0.36	0.37
상위 80%	0.32	0.32	0.32	0.33	0.32	0.33
상위 90%	0.29	0.28	0.29	0.29	0.29	0.29
상위 100%	0.27	0.27	0.27	0.27	0.27	0.27

표 4-6 자차 모형 스코어의 %Response 비교

스코어 구간	Logistic Regression 단일 모형					앙상블 모형
	모형1	모형2	모형3	모형4	모형5	NN3개+Reg2개
상위 10%	31.0	31.0	<b>33.3</b>	33.3	31.0	<b>38.1</b>
상위 20%	50.0	50.0	50.0	47.6	50.0	59.5
상위 30%	71.4	71.4	69.0	71.4	71.4	71.4
상위 40%	73.8	73.8	71.4	76.2	73.8	73.8
상위 50%	88.1	88.1	88.1	88.1	88.1	83.3
상위 60%	95.2	90.5	90.5	90.5	88.1	92.9
상위 70%	95.2	95.2	95.2	97.6	95.2	97.6
상위 80%	97.6	97.6	97.6	97.6	97.6	97.6
상위 90%	100.0	100.0	100.0	100.0	100.0	97.6
상위 100%	100.0	100.0	100.0	100.0	100.0	100.0

표 4-7 자차 모형 스코어의 %Captured Response 비교

스코어 구간	Logistic Regression 단일 모형					앙상블 모형
	모형1	모형2	모형3	모형4	모형5	NN3개+Reg2개
상위 10%	12.0	12.7	<b>14.9</b>	12.1	13.3	<b>12.7</b>
상위 20%	8.9	10.0	9.1	10.0	9.0	10.0
상위 30%	8.0	7.3	7.4	8.0	7.5	8.3
상위 40%	6.6	7.6	6.1	7.5	7.5	7.8
상위 50%	6.2	6.4	6.4	6.2	6.1	6.6
상위 60%	5.8	5.8	5.8	6.0	5.8	6.0
상위 70%	5.1	5.3	5.3	5.4	5.3	5.4
상위 80%	5.0	5.0	4.8	4.9	4.9	4.8
상위 90%	4.5	4.6	4.5	4.6	4.6	4.5
상위 100%	4.3	4.3	4.3	4.3	4.3	4.3

표 4-8 자손 모형 스코어의 %Response 비교

스코어 구간	Logistic Regression 단일 모형					앙상블 모형
	모형1	모형2	모형3	모형4	모형5	NN3개+Reg2개
상위 10%	27.9	30.2	<b>34.9</b>	30.2	32.6	<b>39.5</b>
상위 20%	44.2	46.5	46.5	46.5	44.2	46.5
상위 30%	55.8	51.2	55.8	55.8	53.5	58.1
상위 40%	72.1	74.4	62.8	69.8	69.8	74.4
상위 50%	74.4	76.7	74.4	72.1	76.7	76.7
상위 60%	83.7	81.4	81.4	83.7	81.4	86.0
상위 70%	83.7	86.0	86.0	88.4	86.0	88.4
상위 80%	93.0	93.0	88.4	93.0	93.0	88.4
상위 90%	95.3	95.3	95.3	95.3	95.3	95.3
상위 100%	100.0	100.0	100.0	100.0	100.0	100.0

표 4-9 자손 모형 스코어의 %Captured Response 비교

### 4.3.2 스코어카드 비교

Logistic Regression 단일 모형으로 생성된 스코어카드와 앙상블 모형의 스코어를 이용하여 제안된 방법으로 생성한 스코어카드를 서로 비교해 보았다. 각 변수의 범주별로 부여되는 스코어의 절대적인 크기는 두 가지 모형에서 서로 다르게 나타난다. 그러나 각 변수 내에서 범주별로 부여된 스코어들의 상대적인 크기는 대체적으로 유사하게 나타나므로 앙상블 모형의 스코어를 이용하여 추정된 스코어카드는 Logistic Regression과 동일하지는 않지만 유사한 정도의 설명력을 가질 수 있음을 증명하였다.

피해자의 연령, 사고 접수 후 추가접수 경과기간, 업무용 차량 여부 등의 변수가 사용된 대인 모형의 스코어카드는 다음과 같다.

변수	변수 그룹	그룹별 값 범위	단일모형 스코어	앙상블모형 스코어
Variable 1	1	Variable1 < 16	3	3
	2	16 <= Variable1 < 22	66	49
	3	22 <= Variable1 < 37	71	54
	4	37 <= Variable1 < 65	70	53
	5	65 <= Variable1	64	47
Variable 2	1	Y*>=5	55	60
	2	N*2	56	66
	3	N*>=5	84	93
	4	N*3, N*4, Y*3, Y*4	53	55
	5	N*0, N*1	36	44
	6	Y*1, Y*2	3	3
Variable 3	1	Y*1, Y*2, Y*3, Y*4, Y*>=5	63	64
	2	N*1, N*2, N*3, N*4, N*>=5	26	38
	3	N*0, Y*0	3	3
Variable 4	1	Y*Y	39	44
	2	N*Y	21	27
	3	N*N, Y*N	3	3

표 4-10 대인 모형 스코어카드 비교(1)



변수	변수 그룹	그룹별 값 범위	단일모형 스코어	양상블모형 스코어
Variable 5	1	Z	95	94
	2	ZZ	71	70
	3	X	81	82
	4	N	43	43
	5	Y*2*10, Y*2*20, Y*2*30, Y*2*40, Y*2*>=50, Y*3*10, Y*3*20, Y*3*30, Y*3*40, Y*3*>=50, Y*4*10, Y*4*20, Y*4*30, Y*4*40, Y*4*>=50, Y*>=5*10, Y*>=5*20, Y*>=5*30	3	3
Variable 6	1	Variable6 < -100	3	3
	2	-100 <= Variable6 < 0	38	35
	3	0 <= Variable6 < 4	23	25
	4	4 <= Variable6 < 30	50	55
	5	30 <= Variable6 < 60	71	76
	6	60 <= Variable6	57	71
Variable 7	1	6, 7, 9, A, B	31	31
	2	8	31	29
	3	C	17	16
	4	3, 4, 5	3	5
	5	1, 2	6	3
Variable 8	1	22_N, 22_Y, 99_N, 99_Y	48	41
	2	02_N, 02_Y, 03_N, 03_Y, 07_N, 07_Y, 08_N, 08_Y, 09_N, 09_Y	18	29
	3	01_Y	29	24
	4	12_Y, 19_Y, 20_Y, 21_Y	20	29
	5	04_Y	19	26
	6	01_N	16	17
	7	04_N	22	19
	8	12_N, 19_N, 20_N, 21_N	4	4
Variable 9	1	N*Y	54	59
	2	Y*Y	47	56
	3	N*N	8	8
	4	Y*N	3	3

표 4-11 대인 모형 스코어카드 비교(2)

변수	변수 그룹	그룹별 값 범위	단일모형 스코어	양상블모형 스코어
Variable 10	1	N*N*N*>=9,N*Y*Y*>=9,Y*N*N*>=9	55	46
	2	N*N*N*5~8, N*Y*Y*5~8	38	37
	3	Y*N*N*5~8, Y*Y*Y*5~8	45	46
	4	N* *N*2~4,N*N*N*2~4,N*Y*Y*2~4	36	45
	5	Y*N*N*2~4, Y*Y*Y*2~4	37	45
	6	N**N*Z,N*N*N*Z,N*Y*Y*Z,Y*N*N*Z	3	3
	7	N*N*N*1,N*Y*Y*1,Y*N*N*1,Y*Y*Y*1	19	19
Variable 11	1	Variable11 < 5	18	28
	2	5 <= Variable11 < 13	31	30
	3	13 <= Variable11 < 17	14	13
	4	17 <= Variable11 < 19	23	24
	5	19 <= Variable11 < 22	3	3
	6	22 <= Variable11	22	22
Variable 12	1	Variable12 < -100	52	61
	2	-100 <= Variable12 < -50	3	3
	3	-50 <= Variable12 < 10	12	11
	4	10 <= Variable12 < 90	22	19
	5	90 <= Variable12	11	17
Variable 13	1	Variable13 < -100	171	166
	2	-100 <= Variable13 < -25	7	4
	3	-25 <= Variable13 < 200	6	7
	4	200 <= Variable13	30	31
Variable 14	1	1	9	9
	2	2, 3	27	31
	3	8, 9, ZZ	29	35
	4	7	23	22
	5	12, 13, 14	3	3
	6	4, 5, 6	15	10
Variable 15	1	Variable15 < 1	72	79
	2	1 <= Variable15 < 1000000	3	3
	3	1000000 <= Variable15 < 1500000	32	24
	4	1500000 <= Variable15 < 2000000	31	26
	5	2000000 <= Variable15 < 5000000	39	39
	6	5000000 <= Variable15	71	73
Variable 16	1	Variable16 < 1	47	34
	3	1 <= Variable16 < 1000000	3	3
	4	1000000 <= Variable16 < 1500000	14	13
	5	1500000 <= Variable16	35	28

표 4-12 대인 모형 스코어카드 비교(3)

가해자와 피해자 차량의 동일 정비업체 입고 여부, 기명 피보험자와 차 운전자 불일치 여부, 대인피해자 수 등의 변수가 사용된 대물 모형의 스코어카드를 다음과 같다.

변수	변수 그룹	그룹별 값 범위	단일모형 스코어	앙상블모형 스코어
Variable 1	1	N*0	188	233
	2	Y*3, Y*4, Y*>=5	76	12
	3	N*1, N*2	111	126
	4	Y*2	12	12
	5	N*3, N*4, N*>=5	68	58
	6	Y*0, Y*1	80	68
Variable 2	1	Y	415	334
	2	N	12	12
Variable 3	1	N*Y*Y*2~4, N*Y*Y*5~8, N*Y*Y*>=9	236	403
	2	N* *N*Z, N*N*N*Z, N*Y*Y*Z, Y*N*N*Z, Y*Y*Y*Z	147	12
	3	N*N*N*2~4, N*N*N*5~8, N*N*N*>=9	88	210
	4	N* *N*1, N*N*N*1, N*Y*Y*1	128	139
	5	N*N*N*0, N*Y*Y*0, Y*N*N*0, Y*Y*Y*0	248	359
	6	Y*N*N*1, Y*N*N*2~4, Y*N*N*5~8, Y*N*N*>=9, Y*Y*Y*1, Y*Y*Y*2~4, Y*Y*Y*5~8, Y*Y*Y*>=9	12	141
Variable 4	1	Y*2~4, Y*5~8, Y*>=9	146	27
	2	N*Z, Y*Z	12	12
	3	N*2~4, N*5~8, N*>=9	141	27
	4	N*1, Y*1	12	12
	5	N*0, Y*0	12	12

표 4-13 대물 모형 스코어카드 비교

사고형태, 과거 자차 사고 지급보험금, 과거 가해자 사고건수 등의 변수가 사용된 자차 모형의 스코어카드를 다음과 같다.

변수	변수 그룹	그룹별 값 범위	단일모형 스코어	앙상블모형 스코어
Variable 1	1	Variable1 < -250	199	368
	2	-250 ≤ Variable1 < -100	94	299
	3	-100 ≤ Variable1	8	8
Variable 2	1	Y	248	219
	2	N	8	8
	3	01, 02, 06, 12, 19	237	188
	4	03, 04	197	172
	5	05	106	57
	6	07, 08, 14, 22, 99	8	19
	7	09, 10, 13	41	8
Variable 3	1	Variable3 < 0	114	94
	2	0 ≤ Variable3 < 1	168	181
	3	1 ≤ Variable3 < 2	8	8
	4	2 ≤ Variable3	8	61
Variable 4	1	Variable4 < 1	32	8
	2	1 ≤ Variable4 < 300000	66	19
	3	300000 ≤ Variable4	8	14
Variable 5	1	Variable5 < 1	32	23
	2	1 ≤ Variable5 < 2	8	22
	3	2 ≤ Variable5	80	8

표 4-14 자차 모형 스코어카드 비교

심야사고여부, 최초경찰신고여부, 과거 자손관련 사고 건수 등의 변수가 사용된 자손 모형의 스코어카드는 다음과 같다.

변수	변수 그룹	그룹별 값 범위	단일모형 스코어	양상블모형 스코어
Variable 1	1	02_Y, 03_Y, 04_Y	138	221
	2	02_N, 03_N, 04_N	100	221
	3	06_N, 06_Y, 14_N, 22_N, 22_Y, 99_N, 99_Y	127	199
	4	08_N, 08_Y, 12_N, 12_Y, 19_N, 20_N, 21_N	107	200
	5	01_Y, 05_Y	58	99
	6	07_Y, 09_Y, 10_Y	115	202
	7	07_N, 09_N, 10_N	45	105
	8	01_N, 05_N	10	10
Variable 2	1	N*Y, Y*N	104	104
	2	Y*Y	83	134
	3	N*N	10	10
Variable 3	1	Variable3 < -100	502	255
	2	-100 <= Variable3 < -99	57	47
	3	-99 <= Variable3 < 100	36	23
	4	100 <= Variable3 < 500	10	10
Variable 4	1	Variable4 < 1	93	176
	2	1 <= Variable4 < 500000	10	10
	3	500000 <= Variable4 < 1000000	91	174
	4	1000000 <= Variable4 < 2000000	151	165
	5	2000000 <= Variable5 < 5000000	135	126
	6	5000000 <= Variable5	153	252
Variable 5	1	Variable5 < 1	103	139
	2	1 <= Variable5	10	10

표 4-15 자손 모형 스코어카드 비교

## 제 5장 결론 및 향후 연구과제

일반적으로 보험사기 적발모형에 사용되는 Logistic Regression 단일모형 보다 앙상블 모형의 예측력이 우수하지만 설명력이 약하다는 단점으로 실무에 적용하기에는 어려움이 있었다. 본 연구에서는 보험사기 적발을 위한 입력변수를 생성한 후 데이터마이닝 기법을 활용하여 앙상블 스코어 모형을 생성하였고 앙상블 모형의 예측력이 우수함을 증명하였다. 그리고 설명력을 보완하기 위하여 완성된 앙상블 모형의 스코어를 Scaling하여 모형에 사용된 각각의 변수가 가진 범주별 스코어를 추정할 수 있는 스코어카드를 생성하였고 그 결과는 다음과 같다.

- (1) Logistic Regression 단일모형과 앙상블 모형의 스코어를 상위 10% 구간에서 성능을 비교해본 결과 사기 검출력과 검출효율 측면에서 앙상블 모형의 예측력이 더 우수하였다. 검출력의 경우 앙상블 모형의 사기 검출력이 약 2~5%까지 성능이 향상되는 것을 볼 수 있었다. 데이터 수집의 한계로 샘플데이터 건 수와 변수선택에 제한이 있었던 것을 감안하면 충분한 데이터와 다양한 입력변수가 있다면 성능 향상에서 훨씬 더 높은 효과를 볼 수 있을 것으로 기대된다.
- (2) 앙상블 모형의 설명력을 보완하기 위하여 본 논문에서 제시한 방법으로 스코어카드를 생성해보았다. 그 결과 Logistic Regression 단일 모형과 동일하지는 않지만 유사한 스코어카드를 얻을 수 있었다. 모형에 사



용된 각 변수당 범주별 스코어의 절대적인 크기는 다르지만 변수내에서 범주들이 가지는 상대적인 크기는 유사하게 추정이 가능하였다.

본 연구를 통해서 이와 같은 결과를 얻었으며 보험사기 적발 모형 구축시 설명력 위주가 아니라 사기 적중률에 초점을 맞춘다면 다양한 데이터마ining 기법들을 활용하여 더욱 안정적이고 우수한 성능의 모형을 얻을 수 있음을 확인하였다. 본 논문에서 제시한 방법을 이용한다면 보험사기가 아닌 다른 분야의 모델링에서도 설명력이 약한 모형의 이해를 돕는데 많은 도움이 될 것으로 기대된다.

본 논문에서 연구한 보험사기 적발 모형은 보험료 청구건에 사기로 추정되는 스코어를 부여하는 것으로 각 개인이나 사건에 대한 평가만을 할 수 있고 사건이나 사람들 간의 관련성을 파악하는 것은 불가능하다. 최근 보험사기 추세를 보면 가족단위나 친구모임, 병원이나 보험사직원과 결탁하는 등 집단으로 형성된 보험사기가 급속도로 증가하고 있다. 점차 조직화, 대형화 되어가는 각종 보험사기 범죄를 보다 효율적으로 적발하기 위해서는 조사 대상건을 선별해주는 스코어링 시스템뿐만 아니라 이들 보험사기 공모자들간의 배후관계를 파악하고 관련자들을 적발해낼 수 있는 Link Analysis 기법을 이용한 시스템을 병행하여 운영할 필요가 있을 것이다.



## 참 고 문 헌

- Weisberg, H. I., and R. A. Derrig. "Fraud and Automobile Insurance: A Report on Bodily Injury Liability Claims in Massachusetts," *Journal of Insurance Regulation* 9 (1991), pp.497-541.
- D. Optiz and R. Maclin. "Popular ensemble methods : an empirical study," *Journal of AIR*, Vol.11 (1999), pp.169-198.
- T. G. Dietterich. "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization," *Machine Learning*, Vol.40, No.2 (2000), pp.139-157.
- Byrne, Ciara. "Intelligent Fraud Detection", *Communication*. Vo. 5, No. 2 (2002).
- Stijn Viaene et al. "A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Fraud Detection," *Journal of Risk & Insurance* (2002).
- Siddiqi, Naeem. *Credit Risk Scorecards: developing and implementing intelligent credit scoring*. Hoboken, NJ: John Wiley & Sons, 2006.
- 강성범. <보험사기 방지 및 적발 시스템의 효율적인 구축>, 《비즈니스 인텔리전스 월드》(2003. 12), pp.94~105.
- 김헌수. <비통계적 링크분석을 활용한 보험사기의 효과적 적발방법 연구>, 《보험개발연구》, 제14권, 제2호 (2003. 9), pp.107~146.
- 《금융감독원 보도자료》. 2008-2009.
- 《보험개발원 보도자료》. 2009.





## **ABSTRACT**

### **A Study on Effective Insurance Fraud Detection Model using Data Mining Techniques**

Song, Young Mi

Data Mining Lab.

Department of Industrial Engineering

Graduate School

Hongik University

Insurance fraud, any act committed by a insured with the intent to fraudulently obtain payment, rate has been increasing with recent economic depression. According to the press release, the amount of criminal money and the number of swindlers of insurance fraud are coupled for recent 3 years. Systematic method is needed to expose insurance fraud getting more intelligent.

In this paper, I use various data mining techniques, a way to find statistical rules and patterns using collected data, and research optimal system to expose insurance fraud. First, I apply various data mining techniques like Neural Network and Logistic Regression, then compose the ensemble model for seeking the most effective ways to make better performance to expose insurance fraud.

Score cards, made from scaled scores of model as each categorical rating, are expected that will be able to help to get over difficulty of analysis which is weakness of ensemble model, not only for insurance fraud, but developing other models.

