교차판매 스코어링 모델 개발에 관한 연구

한상태[1] · 이성건[2] · 강현철[3] · 정요천[4]

요 약

기업의 입장에서 가장 중요한 이슈 중 하나는 자사에 있는 많은 고객들 중 회사에 수익을 가져다 줄 수 있는 고객이 누구인가라는 문제이다. 이러한 문제에 대한 기업의 고객 관리 전략 중 하나가 '교차판매(Cross-Selling)' 전략이다.

본 연구에서는 국내 A 손해보험사의 고객 데이터베이스를 활용하여 데이터마이닝 프로세스가 어떻게 진행되고 있는지를 실제 프로젝트를 중심으로 설명하고자 한다. 특히, 본 연구에서 목표로 하고 있는 것은 기존의 자동차보험에 가입한 고객 중에서 장기보험에 추가로 가입하는 고객을 설명하기 위한 교차판매 스코어링 모델을 개발하는 것이다.

주요용어 : 데이터마이닝, 교차판매 스코어링 모델

1. 서론

최근 많은 기업들은 자사가 보유한 고객데이터를 이용하여 시장에서의 경쟁력을 갖출 수 있는 다양한 관점의 모델을 개발하는데 데이터마이닝을 적극 활용하고 있다. 특히, 은행, 카드사, 보험사 등 금융 관련 기업에서 가장 활발히 활용되고 있는데, 이는 금융권 회사들이 고객과의 다양한 접촉을 통해 고객데이터를 확보하는데 심혈을 기울여 왔기 때문이다. 금융 관련 기업들 중 특히, 손해보험 업계에서는 자동차 보험에 대한 이탈모형과 장기보험 상품에 대한 가입모형 개발 등에 큰 관심을 갖고 있다.

이와 관련하여 최근에 한상태 등(2002)은 손해보험사의 자동차보험 가입자에 대한 이탈모형을 개발하여 실제 현업에 활용하고 있다. 본 연구는 한상태 등(2002)에 의한 연구의 확장으로써 손해보험사 자동차보험에 가입한 고객을 대상으로 장기보험상품에 신규가입 가능성을 설명할 수 있는 교차판매 모형을 개발하고자 한다. 이를 통해 자동차보험 갱신율 제고 및 장기보험 상품의 추가 판매율을 향상시켜 기업의 경쟁력을 강화시킬 수 있는 기반을 제공해 주고자 한다. 특히 본 연구는 국내 A 손해보험사에서 실제 진행되었던 데이터마이닝 프로젝트를 중심으로 구성하였는데, 데이터마이닝 소프트웨어로는 SAS사의 Enterprise Miner 4.0을 이용하였다 (SAS, 1997).

2. 모형정의 및 데이터 추출

2.1 모형정의

1) 호서대학교 자연과학부 정보통계학전공 교수, (336-795) 충남 아산시 배방면 세출리 산 29-1

2) 고려대학교 대학원 통계학과 박사과정, (136-701) 서울 성북구 안암동 5가 1번지

3) 호서대학교 자연과학부 정보통계학전공 교수, (336-795) 충남 아산시 배방면 세출리 산 29-1

4) 호서대학교 대학원 수학과 통계전공 석사과정, (336-795) 충남 아산시 배방면 세출리 산 29-1

자동차/장기보험 교차판매 모형은 자동차보험만 가입하고 장기보험 상품에는 가입하지 않은 고객을 대상으로 장기보험 상품의 신규가입확률이 높은 고객리스트를 산출하여 추가가입을 유도하고자 하는 것이 목적이었다. 이에 따른 모형의 목표변수는 2000년 1월 31일에서 2001년 1월 31일 사이에 자동차보험에 가입한 고객 중에서 장기보험 상품에의 추가가입 여부이며, 설명변수는 자동차보험 계약 건에 대한 보험 시기/종기 기준의 고객 속성 정보 및 거래 정보를 이용하였다. 또한 마케팅 부서와의 협의를 통해 장기보험의 성격에 따라 가입패턴이 다르다고 판단되어 1차적으로 장기보험에의 가입여부 및 세부적으로 장기보험 상품별로 "상해보험", "질병보험", "암보험", "화재보험"으로 구분하여 총 5개의 모형을 개발하였다. 따라서 먼저 장기보험 상품에 추가 가입 가능성을 살펴보고, 장기보험 상품에 가입 가능성이 높은 고객의 경우 세부적으로 장기보험 상품 중 어느 상품에 가입 가능성이 가장 높은가를 알아보고자 하였다.

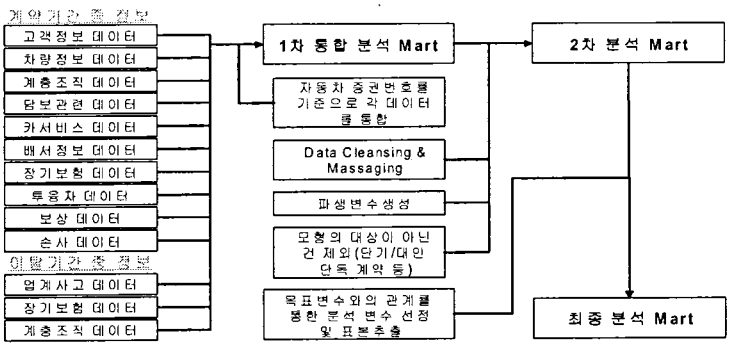
2.2 데이터 추출

앞서 정의한 모형에 필요한 데이터를 A 손해보험사의 기간계 DB와 DW에서 자동차보험 2000년 1월 31일 기준으로 최근 5년 혹은 3년까지의 고객 속성정보 및 거래정보를 추출하였고, 또한 2000년 1월 31일에서 2001년 1월 31일 사이의 개인 고객의 장기보험 상품 가입여부 정보를 추출하였다.

3. 분석용 마트 구성

장기보험 상품에 가입 가능성이 높은 고객을 예측하는데 필요하다고 판단되는 각 데이터를 A 손해보험사의 기간계 시스템에서 추출한 후, 자동차 증권번호를 기준으로 통합하여 1차 통합 분석 마트를 구성하였다. 다음으로 데이터 정제와 파생변수 생성, 모형의 대상이 아닌 변수 제거 등을 통해 2차 분석 마트를 구성하였다. 마지막으로 목표변수와의 관계를 통한 분석변수 선정 및 표본 추출을 통하여 최종 분석 마트를 구성하였다.

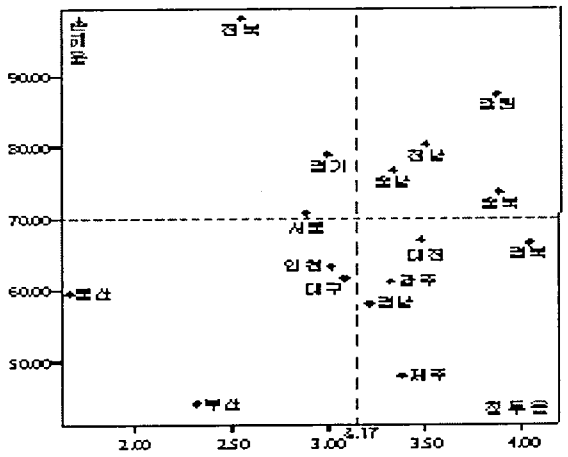
<표 3-1> 분석용 마트의 구성 흐름도



3.1 데이터 탐색

추출된 초기 데이터는 입력오류, 결측값 등이 포함되어 있는 경우가 많으므로 좋은 모형을 개발하기 위해서는 이에 대한 탐색과 정제가 이루어져야 한다. 이를 위해 기초통계분석 등을 통하여 데이터의 충실도, 변수간의 상충, 여러 의미 있는 정보를 살펴보았다. 예를 들어, 권역별 -손해율/침투율 분석을 해 본 결과 울산과 부산지역은 손해율이 낮아서 수익이 클 것으로 예상되는 지역임에도 불구하고 침투율이 낮은 문제를 안고 있음을 발견할 수 있었다. 따라서 기업 입장에서 보다 공격적인 판매전략을 세워야 한다는 것을 알 수 있을 것이다(그림 3-1 참조).

<그림 3-1> Positioning Map(권역) - 손해율/침투율(실적)



3.2 데이터 정제 및 파생변수 생성

고객의 장기보험 상품 추가가입에 관련된 변수들에 포함되어 있는 결측값(Missing Value) 및 오류를 파악한 후 제거 또는 적절한 값으로 수정 변환하는 과정을 거쳤다. 먼저 1단계는 결측값(Missing Value) 및 잡음(Noise) 데이터의 비율이 90% 이상이 되는 필드들을 제거하고, 2단계는 필드간 상충(예 : 보험 개시일자가 만기일자 보다 이후인 경우), 업무적인 결측값 처리 등은 변수간 교차 검증(Cross Check)을 통해 적절한 값으로 변환하였다. 3단계는 입력오류(예 : 나이 필드에 문자가 입력된 경우), 값 범위 초과 및 결측값 등은 해당 레코드를 삭제하였다. 또한, 인수된 데이터 이외에 목표변수에 유의한 영향을 준다고 생각하는 변수를 추가적으로 생성하였는데, 이는 현업과 충분한 협의를 통해 진행되었다.

3.3 표본추출

표본추출을 한 이유는 추가가입자와 비가입자의 비율 차이가 너무 커서 가입자의 특성이 모형에 잘 나타나지 않아 모형개발의 어려움이 있어 추가 가입한 고객은 전체를 추출하고, 비가입한 고객은 가입한 고객의 3배를 표본추출 하였다. 또한 세부모형의 경우 현업의 여건상 장기보험에 추가가입하고 각 세부 모형별 만기월 기준 보험 유지여부를 목표변수로 하여 최종 분석용 마트를 구성하였다(표 3-1 참조).

<표 3-1> 최종 분석용 마트

모형 종류	필드수	전체건수	가입건수	비가입건수	가입율	비가입율
장기추가가입모형	307	53,860	13,465	40,395	25%	75%
상해보험 모형	343	24,768	6,192	18,576		
질병보험 모형	322	5,004	1,251	3,753		
암보험 모형	317	3,656	914	2,742		
화재보험 모형	357	4,868	1,217	3,651		

3.4 변수선택

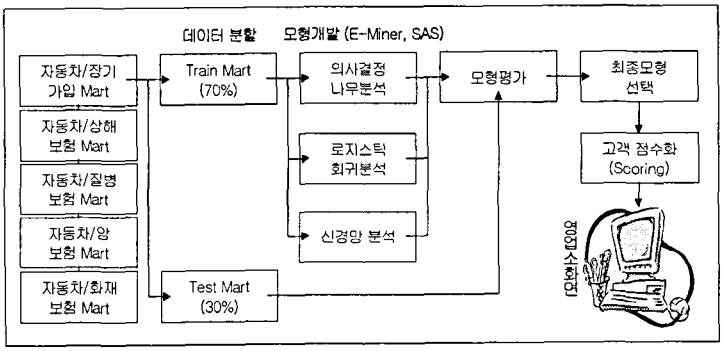
분석변수의 선정 단계에서는 실무자와의 협의 및 목표변수와의 관계분석을 통해 각 모형별로 입력변수를 선정하였는데, 목표변수를 예측하는데 연관성이 낮은 변수를 제거하고자 범주형 변수인 경우에는 카이제곱(Chi-Square)통계량을 연속형 변수인 경우에는 분산분석을 이용하였다.

4. 교차판매 스코어 모형 개발

4.1 분석흐름도

모형 개발의 분석 흐름도는 <그림 4-1>과 같이 구성하였다. 모형개발을 위해 각 모형별 최종 분석용 마트를 훈련용 부분(Train Part) 70%, 평가용 부분(Test Part) 30%로 분할한 후 의사결정나무모형, 로지스틱 회귀모형 및 신경망모형을 수행한 후, 최종적으로 각 모형을 비교 평가하여 최적의 모형을 선택하였다(강현철·한상태 외, 2001).

<그림 4-1> 모형개발 분석흐름도



4.2 의사결정나무분석 결과

자동차/암보험 마트에 대해 의사결정나무분석을 실시한 결과는 다음과 같다.

<표 4-1> 의사결정나무에 의한 암보험 추가가입 모형의 오분류표

예측 / 실제	학습 결과			테스트 결과		
	미가입	가 입	합 계	미가입	가 입	합 계
미가입	1,859	50	1,909	808	25	833
가 입	387	263	650	161	103	264
	Prior Distribution : 75.0% 학습 정분류율 : 82.9%			Prior Distribution : 75.0% 테스트 정분류율 : 83.1%		

<표 4-1>의 결과를 살펴보면, 암보험 추가가입모형에 대해 학습 결과의 정분류율은 82.9%이고, 테스트 결과의 정분류율은 83.1%로서 안정적인 모형임을 알 수 있다. 이 모형에 대한 추가가입 규칙의 일부를 살펴보면 다음과 같다.

① 만기일 기준 대출여부가 있고, 자기손해 가입금액이 20,000원 이하이면 암보험 추가가입 가능성이 94.8%이다.

② 자필서명 구분ID가 자필서명아님, 미분류이고, 계약자와 집금자의 나이 차이가 4.5세미만이고, 1년전 무보험 가입유무가 있고, 만기일 기준 대출여부가 있고, 자기손해 가입금액이 20,000원 이하이면 암보험 추가가입 가능성이 93.1%이다.

4.3 로지스틱 회귀분석 결과

질병보험 추가가입모형에 대해 로지스틱 회귀분석을 수행한 결과는 다음과 같다.

<표 4-2> 로지스틱 회귀분석에 의한 질병보험 추가가입모형의 오분류표

예측 / 실제	학습 결과			테스트 결과		
	미가입	가 입	합 계	미가입	가 입	합 계
미가입	2,468	163	2,631	1,049	73	1,122
가 입	387	485	872	165	214	379
	Prior Distribution : 75.0% 학습 정분류율 : 84.3%			Prior Distribution : 75.0% 학습 정분류율 : 84.2%		

4.4 신경망분석 결과

상해보험 추가가입모형에 대해 신경망분석을 수행한 결과는 다음과 같다.

<표 4-3> 신경망분석에 의한 상해보험 추가가입모형의 오분류표

예측 / 실제	학습 결과			테스트 결과		
	미가입	가 입	합 계	미가입	가 입	합 계
미가입	12,045	921	12,966	5,110	500	5,610
가 입	2,491	1,881	4,372	1,078	742	1,820
	Prior Distribution : 75.0% 학습 정분류율 : 80.3%			Prior Distribution : 75.0% 학습 정분류율 : 78.8%		

4.5 모형평가 및 최종모형 선택

5개 모형에 대한 학습 결과와 테스트 결과는 <표 4-4>와 같다. 결과를 살펴보면 학습결과의 정분류율은 장기상품 추가가입과 상해보험 추가가입 모형에서는 로지스틱 회귀모형이 다소 우세하고, 나머지 모형에 대해서는 신경망모형이 약간 더 우수한 것을 알 수 있다. 그러나 모형의 안정성과 시스템 이식성 및 향후 현장 활용 용이성을 고려하여 로지스틱 회귀모형을 교차판매 스코어링에 사용할 최종모형으로 선택하였다.

<표 4-4> 각 모형에 대한 비교

모형의 정분류율 비교		학습 결과	테스트 결과
의사결정 나무모형	장기상품 추가가입	77.0%	75.7%
	상해보험 추가가입	78.9%	78.2%
	질병보험 추가가입	80.9%	80.6%
	암보험 추가가입	82.9%	83.0%
	화재보험 추가가입	78.0%	77.9%
로지스틱 회귀모형	장기상품 추가가입	78.2%	77.0%
	상해보험 추가가입	80.4%	79.6%
	질병보험 추가가입	84.3%	84.2%
	암보험 추가가입	83.5%	83.0%
	화재보험 추가가입	82.2%	78.9%
신경망 모형	장기상품 추가가입	78.0%	76.9%
	상해보험 추가가입	80.3%	78.8%
	질병보험 추가가입	85.4%	82.9%
	암보험 추가가입	87.0%	82.0%
	화재보험 추가가입	82.6%	78.9%

4.6 고객 스코어링

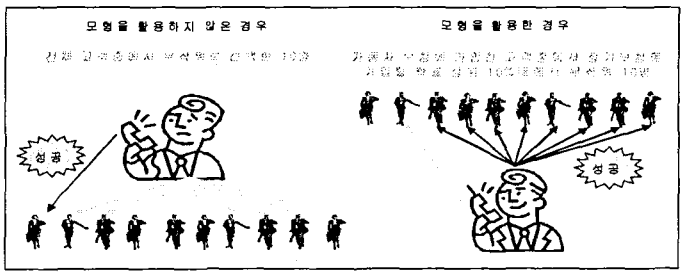
개발된 모형에 대해 장기보험 세부 상품들의 고객별 가입 가능 스코어를 <표 4-5>와 같은 형태로 산출하였다. 이러한 결과를 바탕으로 상대적으로 가입 가능성이 높은 고객을 대상으로 추천상품을 선정하여 마케팅 활동을 전개하고자 하는 것이다. 예를 들어 한 고객이 <표 4-5>와 같은 결과를 얻었다면 이 고객은 장기보험 전체 가입 확률이 0.834로 매우 높고, 추천 세부 상품으로는 가장 높은 가입 확률을 보이는 질병보험 상품을 추천할 수 있는 것이다.

<표 4-5> 상품별 추가가입 확률의 예

증권번호	장기가입확률	상해보험	질병보험	암보험	화재보험
4001****7232****0000	0.874	0.149	0.763	0.598	0.017

4.7 개발된 모형을 활용한 캠페인 활동

<그림 4-3> 모형을 통한 캠페인 활동의 효율성



일반적으로 모형을 고려하지 않을 경우, 기존고객이나 외부 리스트에 수록된 잠재고객들을 대상으로 무작위로 고객을 선정하여 보험 상품의 판매 활동에 대한 캠페인 활동(보험설계사, DM, TM 등)을 벌일 것이다. 그러나 모형을 통한 캠페인 활동을 할 경우, 무작위로 고객을 선정하여 판매 활동을 실시하는 것보다 반응확률이 높을 것으로 예상되는 고객의 리스트를 추출하여 이들을 대상으로 캠페인 활동을 전개함으로서 마케팅비용의 절감과 설계사의 활동 효율성을 제고할 수 있을 것이다.

5. 결론

본 연구의 목표는 자동차/장기보험 교차판매 모형을 개발하는 것이었는데, 로지스틱 회귀모형이 다른 모형에 비해 안정적인 결과를 제공한다는 것을 알 수 있었다. 개발된 모형을 근거로 캠페인 활동(Campaign Management)을 실시하게 되면 마케팅의 효율성을 증대시켜 회사의 이익을 극대화 할 수 있을 것이라 판단된다.

참고문헌

[1] 강현철 · 한상태 · 최종후 · 김은석 · 김미경(2001), 『SAS Enterprise Miner를 이용한 데이터 마이닝 -방법론 및 활용-』, 서울 : 자유아카데미.

[2] 한상태 · 이성건 · 강현철 · 유동균(2002), Development of Scoring Model on Customer Attrition Probability by Using Data Mining Techniques, *The Korean Communication in Statistics Vol.9*, 271-280.

[3] SAS Institute. (1997), *Data Mining Using SAS Enterprise Miner Software*, SAS Institute Inc.