응용통계연구 제 17권 2호, 2004년, pp. 229-238

교차판매(CROSS-SELL) 스코어링 모형 개발

한상태[1] 강현철[2] 이성건[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-sell)' 전략이다. 본 연구에서는 국내 A 손해보험사의 고객 데이터베이스를 활용하여 데이터마이닝 모형 개발이 어떻게 진행되고 있는지 실제 프로젝트를 중심으로 설명하고자 한다. 특히, 본 연구에서 목표로 하고 있는 것은 기존의 자동차보험에 가입한 보험사 고객 중에서 장기보험 및 세부 보험(상해, 질병, 암, 화재 보험 등)에 추가로 가입하는 고객의 특성을 파악하기 위한 교차판매 스코어링 모형을 개발하는 것이다.

주요용어: 데이터마이닝, 교차판매, 스코어링 모형

1. 서론

최근 많은 기업들은 자사가 보유한 고객 데이터를 이용하여 시장에서의 경쟁력을 갖출 수 있는 다양한 관점의 모형을 개발하기 위해 데이터마이닝을 적극 활용하고 있다. 특히, 은행, 카드사, 보험사 등 금융 관련 기업에서 가장 활발히 활용되고 있는데, 이는 금융권 회사들이 고객과의 다양한 접촉을 통해 고객데이터를 확보하는 데 심혈을 기울여 왔기 때문이다. 금융 관련 기업들 중 특히 손해보험 업계에서는 자동차 보험에 대한 이탈모형과 장기보험 상품에 대한 가입모형 개발 등에 가장 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 확장으로서 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 SAS 사의 Enterprise Miner V4를 이용하였다(SAS Institute, 2000, 2003).

1) (336-795) 충청남도 아산시 배방면 세출리, 호서대학교 자연과학부 정보통계학전공, 부교수
E-mail: sthan@office.hoseo.ac.kr
2) (336-795) 충청남도 아산시 배방면 세출리, 호서대학교 자연과학부 정보통계학전공, 전임강사
E-mail: hychkang@office.hoseo.ac.kr
3) (136-701) 서울시 성북구 안암동 5가, 고려대학교 대학원 통계학과, 박사과정
E-mail: sklee@korea.hoseo.ac.kr
4) (336-795) 충청남도 아산시 배방면 세출리, 호서대학교 대학원 통계전공, 석사졸업
E-mail: yocjung@hotmail.com

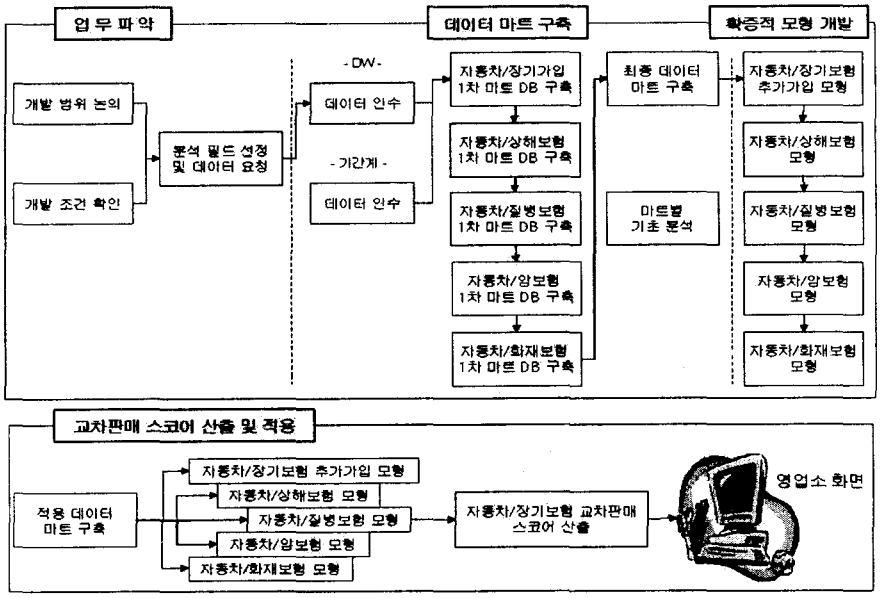

그림 2.1: 프로젝트의 진행 순서

2. 프로젝트 일정 및 데이터마이닝 방법론

본 프로젝트의 목적은 A 손해보험사 자동차보험에 가입한 고객 중 장기보험 상품에 신규 가입하는 고객을 설명할 수 있는 모형을 개발하여 회사의 경쟁력을 강화시키는 것이라 할 수 있다. 이러한 목적 하에 프로젝트의 진행 순서는 그림 2.1과 같이 추진되었다. 먼저 분석자들이 보험업무 및 운영계/정보계 데이터의 구조를 이해할 수 있도록 업무파악 과정이 진행되었고, 이를 기초로 필요한 사내외 데이터들이 선정되어 해당 데이터 담당부서에 요청되었다. 또한 전달된 데이터들을 가공하고 통합하여 분석용 데이터 마트(data mart)가 구성되었으며, 이를 이용하여 상품별 스코어링 모형이 개발되었다. 마지막으로 개발된 스코어링 모형에 근거하여 고객별 교차판매 스코어(점수)가 계산되었고, 스코어가 높은 고객들을 선정하여 캠페인이 진행되었다.

또한, 데이터마이닝 모형개발 프로젝트에는 CRISP-DM(cross industry standard process of data mining) 방법론이 적용되었으며(Chapman et al., 1999; Chapman et al., 2000), 본 프로젝트의 진행 순서는 비즈니스 이해, 데이터 이해, 데이터 준비, 모형 구축, 평가, 전개 등 여섯 단계로 하였다. 첫번째 단계인 비즈니스 이해단계는 데이터마이닝 목표 결정 및 프로젝트 계획을 수립하며, 두 번째 단계인 데이터 이해는 데이터 기술, 탐색, 데이터 평가를 시행하고, 세 번째 단계인 데이터 준비는 데이터의 설정, 선택, 정제, 생성, 통합을 한다. 네 번째 모형 구축단계에서는 모형 기법을 선택, 테스트 설계와 모형 생성, 평가를 시행하며, 다섯 번째 평가단계에서는 모형을 평가하고, 마지막 여섯 번째 단계인 전개 단계에서는 최종보고 및 마케팅 캠페인 실행을 진행하였다(그림 2.2 참조).

비즈니스 이해	데이터 이해	데이터 준비	모형 구축	평 가	전 개
업무목적 결정 -배경 -업무 목적 -업무 성공기준 **상황평가** -자원 관리 -요구,가정,제약 -위험과 상황 -용어 -비용과 효용 **DM 목표 결정** -DM 목표 -DM 성공기준 **계획 수립** -프로젝트 계획 -툴,기법의 평가	**데이터 수집** -수집 보고서 **데이터 기술** -데이터 기술서 **데이터 탐색** -탐색 결과 보고서 **데이터 평가** -평가 보고서	**데이터 설정** -데이터 기술서 **데이터 선택** -포함제외 사유서 **데이터 정제** -정제 보고서 **분석용 데이터 마트 구축** -변수 변환 -레코드 생성 **데이터 결합** -통합 데이터 **데이터 포맷** -그룹변수 생성 -재구성 포맷	**모형 기법 선택** -모형 기법 -모형 설정 가정 **테스트 구조** -테스트 실행결과 **모형 구축** -결정된 모수 기술 -구축 모형 -모형 기술서 **모형 평가** -모형 평가서 -변경된 모수 기술	**결과 평가** -업무 성공기준에 의한 DM의 결과 평가 -최종 모형 기술 **작업과정 재검토** -검토결과 보고서 **대안 제시** -가능한 여러 대안에 대한 기술서	**전개계획 수립** -전개 계획서 **모니터링 및 유지계획 수립** -모니터링 계획서 -유지 계획서 **최종보고** -최종 보고서 -최종보고 **프로젝트 전반에 관한 검토** -검토 보고서

그림 2.2: 단계별 주요 작업 및 결과물

3. 모형 정의 및 데이터 추출

3.1. 모형 정의

본 연구에서 개발하고자 하는 자동차/장기보험 교차판매 모형은 자동차보험만 가입하고 장기보험 상품에는 가입하지 않은 고객을 대상으로 장기보험 상품의 신규가입확률이 높은 고객리스트를 산출하여 추가가입을 유도하고자 하는 것을 목적으로 추진되었다. 이에 따른 모형의 목표변수는 2000년 1월 31일에서 2001년 1월 31일 사이에 자동차보험에 가입한 고객 중에서 장기보험 상품에의 추가가입 여부이며, 자동차보험 가입 시점의 고객, 차량, 조직, 담보, 서비스, 배서, 장기보험, 투융자, 보상, 손사 정보 등 약 400여개의 변수가 설명변수로 수집되었다.

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.

3.2. 데이터 추출

모형 개발에 필요한 데이터들은 A 손해보험사의 기간계 데이터베이스(database)와 데이터웨어하우스(data warehouse)로부터 수집되었는데, 자동차보험 가입 시점의 고객속성

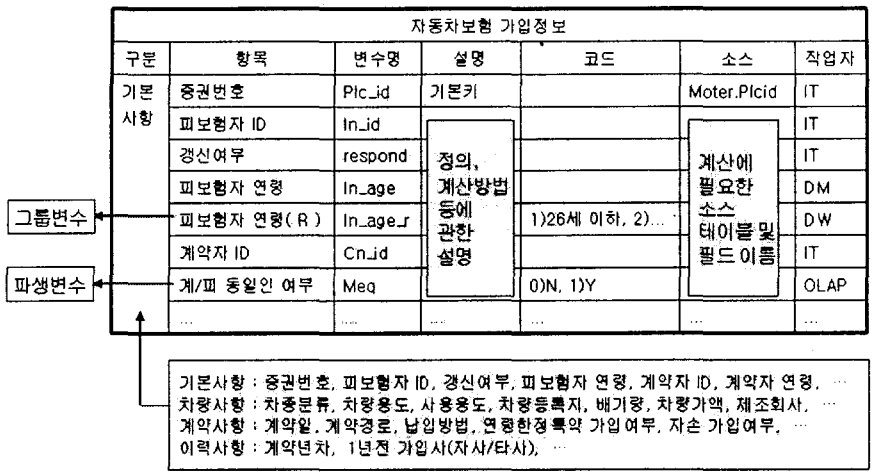

그림 3.1: 자동차보험 가입 정보 수집방안의 설계

정보(성, 나이, 위험등급 등) 및 자동차보험 가입 정보(차량종류, 배기량, 한정특약 가입여부 등; 그림 3.1 참조) 그리고 최근 5년 혹은 3년까지의 기타 거래정보(최근 1년간, 2년간, ···, 5년간 보상회수 등)가 추출되었다. 또한 2000년 1월 31일에서 2001년 1월 31일 사이의 개인 고객의 장기보험 상품 가입여부 정보를 추출하였다.

4. 분석용 데이터마트의 구성

장기보험 상품에 가입 가능성이 높은 고객을 예측하는 데 필요하다고 판단되는 각 데이터를 A 손해보험사의 기간계 시스템에서 추출한 후, 자동차 증권번호를 기준으로 통합하여 1차 분석용 데이터마트(data mart)를 구성하였다. 다음으로 데이터 정제와 파생변수 생성, 모형의 대상이 아닌 변수 제거 등을 통해 2차 분석용 마트를 구성하였다. 마지막으로 목표변수와의 관계를 통한 분석변수 선정 및 표본추출을 통하여 최종 분석용 마트를 구성하였다(그림 4.1 참조).

4.1. 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 통해 여러 가지 의미 있는 정보를 살펴보았으며, 데이터의 충실도(기록율)를 조사하여 사용 불가능한 필드들을 파악하였다.

특히, 주소변경회수, 개인근무지, 직위, 학력 등 고객의 인구사회적 속성과 관련된 정보들은 모형개발에 있어서 중요한 역할을 할 수 있음에도 불구하고 기록율이 매우 떨어져 입력변수로 사용할 수 없었다. 따라서 향후 캠페인을 통한 고객정보 획득 혹은 외부데이터를

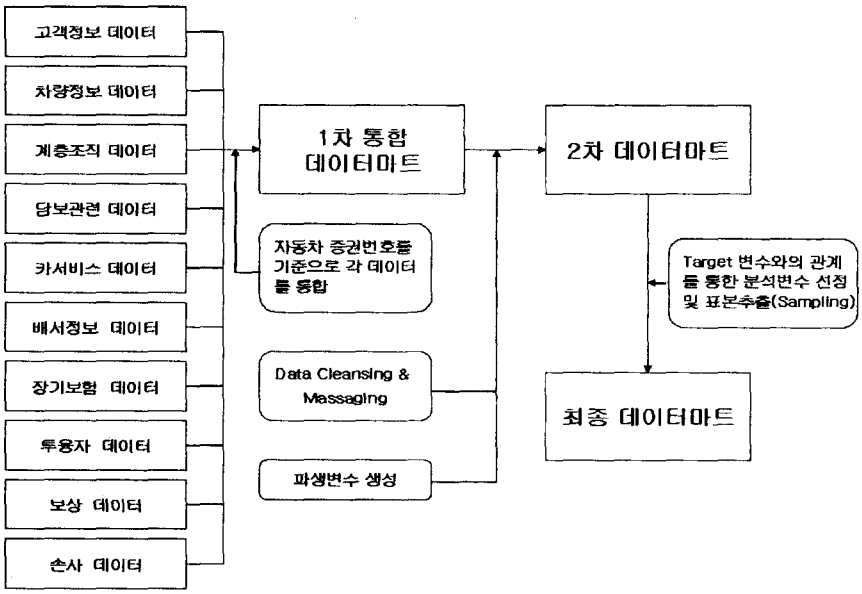

그림 4.1: 분석용 데이터마트 구성의 흐름도

활용한 데이터 보강이 이루어진다면 보다 안정적인 스코어링 모형을 개발할 수 있을 것으로 생각된다.

4.2. 데이터 정제 및 파생변수 생성

고객의 장기보험 상품 추가가입에 관련된 변수들에 포함되어 있는 결측값 및 오류를 파악하여 제거하거나 적절한 값으로 수정 변환하는 과정을 수행하였다. 먼저 1단계에서는 결측값 및 오류 데이터의 비율이 50% 이상이 되는 필드들을 제거하였고, 2단계에서는 필드간 상충(예를 들어, 보험 개시일자가 만기일자 보다 이후인 경우) 또는 업무적인 결측값이 발견된 경우 변수들 간의 교차분석을 통해 적절한 값으로 변환하였다. 3단계에서는 입력오류(예를 들어, 나이에 문자가 입력된 경우), 범위 초과, 결측값 등에 대해 해당 레코드를 삭제하였다. 또한, 인수된 데이터 이외에 목표변수에 유의한 영향을 준다고 생각하는 변수를 추가적으로 생성하였는데 이는 현장의 실무자들과 충분한 협의를 통해 진행되었다.

4.3. 표본추출

추가가입자와 미가입자의 비율 차이가 너무 커서 가입자의 특성이 모형에 잘 나타나지 않아 모형개발이 곤란한 경우가 있기 때문에, 추가가입한 고객은 전체를 추출하고 비가입한 고객은 일부를 표본추출하였다. 본 연구에서는 여러 가지 추출비율에 대하여 표본추출이 검토되었으며, 분석에 소요되는 시간 등을 고려하여 장기보험 전체의 표본크기가 대략 50,000 정도가 되도록 하기 위해 최종적으로는 미가입한 고객이 가입한 고객의 3배가 되도

표 4.1: 최종 분석용 데이터마트의 표본 현황

세부 모형	필드 수	전체건수	가입건수	미가입건수	가입율	미가입율
장기보험 전체	307	53,860	13,465	40,395	25%	75%
상해보험	343	24,768	6,192	18,576		
질병보험	322	5,004	1,251	3,753		
암보험	317	3,656	914	2,742		
화재보험	357	4,868	1,217	3,651		

록 하였다. 또한 세부모형의 경우 만기월 기준 자동차보험을 유지한 상태에서 해당 장기보험에 추가가입을 하였는지의 여부를 목표변수로 하여 최종 분석용 데이터마트를 구성하였다(표 4.1 참조).

4.4. 변수선택

분석변수의 선정 단계에서는 실무자와의 협의 및 목표변수와의 관계분석을 통해 각 모형별로 입력변수를 선택하였는데, 목표변수를 예측하는데 연관성이 지나치게 낮은 변수를 미리 제거하고자 범주형 변수인 경우에는 카이제곱검정을 그리고 연속형 변수인 경우에는 분산분석을 이용하였다. 본 연구에서는 유의확률이 0.5 이상인 변수들을 대상으로 실무자와의 협의를 거쳐 최종적으로 제거할 변수를 선정하였다. 이 단계에서 제거된 설명변수들은 대략 전체의 30% 정도이다.

5. 교차판매 스코어링 모형 개발

5.1. 분석흐름도 및 결과

모형 개발의 분석 흐름도는 그림 5.1과 같이 구성하였다. 모형개발을 위해 각 세부모형별 최종 분석용 데이터마트를 학습용(training data) 70%, 평가용(test data) 30%로 분할하여 의사결정나무분석, 로지스틱 회귀분석, 신경망분석 등 세 가지 기법을 적용한 후 최종적으로 각 기법의 결과들을 비교 평가하여 최적의 모형을 선택하였다(이들 기법에 대해서는 강현철 등(2001)을 참조하기 바란다).

기법 적용 결과의 한 예로서 자동차/암보험 데이터마트에 대해 의사결정나무분석을 실시한 결과는 표 5.1과 같다(이 결과는 분류기준값(threshold value)을 0.5로 하여 계산된 것이다). 이를 살펴보면 암보험 추가가입 모형에 대해 학습용 데이터에 대한 정분류율은 83.8%이고 평가용 데이터에 대한 정분류율은 83.0%로서 비교적 안정적인 결과를 보여주고 있다.

또한 의사결정나무분석에 의한 추가가입 규칙의 일부를 제시하면 다음과 같다; (1) 만기일 기준 대출금액이 있고, 자기손해 가입금액이 20,000원 이하이면 암보험 추가가입 가능성이 94.8%이다. (2) 자필서명 구분 ID가 자필서명 아님 또는 미분류이고, 계약자와 집금

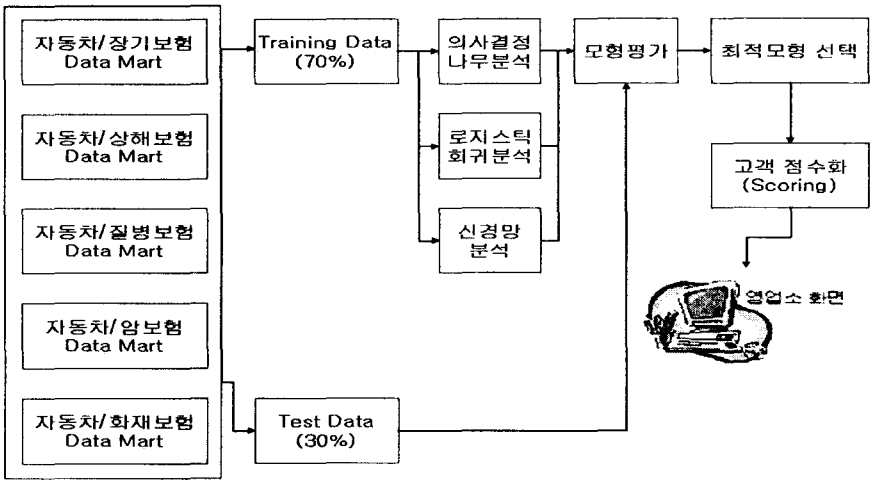

그림 5.1: 모형개발 분석흐름도

표 5.1: 의사결정나무에 의한 암보험 추가가입 모형의 정오분류표

예측 / 실제	학습용 데이터			평가용 데이터		
	미가입	가 입	합 계	미가입	가 입	합 계
미가입	1,859	50	1,909	808	25	833
가 입	363	287	650	161	103	264
	정분류율 : 83.8%			정분류율 : 83.0%		

자의 나이 차이가 4.5세 미만이고, 1년전 무보험 가입경험이 있고, 만기일 기준 대출금액이 있고, 자기손해 가입금액이 20,000원 이하이면 암보험 추가가입 가능성이 93.1%이다.

5.2. 모형평가 및 고객 스코어링

표 5.2는 학습용 데이터와 평가용 데이터에 각 기법을 적용하여 얻은 정분류율을 정리한 것이다. 이 결과는 분류기준값(threshold value)을 0.5로 하여 계산된 것이며, 일반적으로는 베이즈 규칙(Bayes rule)을 사용하여 오분류율을 최소로 하거나 이익(profit)을 최대로 하는 분류기준값을 정할 수 있다(Breiman et al., 1984; Ripley, 1996). 또한 표본에서의 목표변수의 각 범주별 비율(ρ_0, ρ_1)이 사전확률(π_0, π_1)과 다른 경우, 모집단과 표본의 이러한 비율차이를 보정한 사후확률의 추정치 $\hat{p} = \hat{P}(Y = 1|x)$는 다음과 같이 계산될 수 있다(SAS Institute, 2002).

$$\hat{p} = \frac{\hat{p}^* \rho_0 \pi_1}{(1 - \hat{p}^*)\rho_1 \pi_0 + \hat{p}^* \rho_0 \pi_1}$$

여기서 $\hat{p}^*$는 보정되지 않은 사후확률의 추정치이다(본 연구에서는 $\hat{p}^*$를 사용하였다).

표 5.2: 세부 모형에 대한 정분류율

기법	세부 모형	학습용 데이터	평가용 데이터
의사결정 나무분석	장기보험 전체	77.0%	75.7%
	상해보험	78.9%	78.2%
	질병보험	80.9%	80.6%
	암보험	83.8%	83.0%
	화재보험	78.0%	77.9%
로지스틱 회귀분석	장기보험 전체	78.2%	77.0%
	상해보험	80.4%	79.6%
	질병보험	84.3%	84.2%
	암보험	83.5%	83.0%
	화재보험	82.2%	78.9%
신경망 분석	장기보험 전체	78.0%	76.9%
	상해보험	80.3%	78.8%
	질병보험	85.4%	82.9%
	암보험	87.0%	82.0%
	화재보험	82.6%	78.9%

표 5.3: 상품별 추가가입 스코어링의 예

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

표 5.2를 살펴보면 학습용 데이터에 대한 정분류율은 장기보험 전체와 상해보험에 대해서는 로지스틱 회귀분석이 다소 높고, 다른 세부 보험상품에 대해서는 신경망분석이 약간 더 높은 것을 알 수 있다. 또한 평가용 데이터에 정분류율은 전반적으로 로지스틱 회귀분석이 높음을 알 수 있다. 따라서 모형의 안정성 및 캠페인 시스템 개발에 따른 프로그램 작성의 용이성 등을 이유로 로지스틱 회귀분석의 결과를 이용하여 교차판매 스코어링에 사용할 최종모형을 개발하였다.

한편, 개발된 최종모형에 의해 장기보험 세부 상품들의 고객별 추가가입 스코어를 표 5.3과 같은 형태로 산출하였으며, 이러한 결과를 바탕으로 상대적으로 추가가입 가능성이 높은 고객을 대상으로 추천상품을 선정하여 마케팅 활동을 전개하였다. 예를 들어 한 고객에 대해서 표 5.3과 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 있을 것이다. 이와 같이 반응확률이 높을 것으로 예상되는 고객들을 대상으로 켐페인 활동을 전개함으로써 마케팅 비용의 절감과 설계사의 활동 효율성을 제고할 수 있다.

실제 프로젝트에서는 추가가입 가능성이 있는 고객이 캠페인 대상에 충분히 포함될 수 있는 분류기준값들을 탐색한 후, 실무자와의 협의를 거쳐 대략 전체고객의 30% 정도를 캠

페인 대상으로 하였다(즉, 전체 고객에 대해 사후확률을 계산하여 내림차순으로 정렬한 후 상위 30%를 캠페인 대상으로 선정하였다).

6. 결론

본 연구에서는 자동차 보험에 가입하고 있는 고객을 대상으로 그들을 우수고객화 하기 위해 장기보험에 추가 가입시키는 자동차/장기보험 교차판매 스코어링 모형을 개발하였으며, 이 모형을 근거로 캠페인 활동을 실시함으로써 마케팅의 효율성을 증대시켜 회사의 이익을 극대화 할 수 있을 것이라 판단된다. 또한 기업의 실제 데이터마이닝 프로젝트에서 적용되었던 분석사례를 통해 프로젝트의 진행순서, 모형정의 및 데이터 추출 과정, 분석용 데이터마트의 생성 과정, 고객점수의 산출 및 적용 과정 등에 대해 하나의 전형을 제시하고자 하였고, 이러한 내용은 유사한 프로젝트를 진행하고자 하는 여러 분야의 독자들이 보다 효율적으로 작업을 수행할 수 있도록 도움을 줄 수 있을 것이다.

한편, 본 연구에서 개발한 교차판매 스코어링 모형은 각 업종의 특성에 맞게 보완함으로써 유사한 관련 산업의 교차판매 모형으로도 활용될 수 있을 것으로 생각된다. 다만, 향후 실제 마케팅 활동으로부터 얻어지는 반응분석 등을 통해서 지속적으로 모형의 성능을 향상시킬 필요가 있으며, 최종적으로 이러한 프로세스를 자동화 할 수 있는 시스템 개발이 필요하다고 여겨진다.

참고문헌

강현철, 한상태, 최종후, 김은석, 김미경 (2001). <SAS Enterprise Miner를 이용한 데이터마이닝 – 방법론 및 활용 –>, 서울 : 자유아카데미.

한상태, 이성건, 강현철, 유동균 (2002). Development of scoring model on customer attrition probability by using data mining techniques, <한국통계학회 논문집>, **9**, 271-280.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C.J. (1984). *Classification and Regression Tree*, Chapman and Hall.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). *CRISP-DM 1.0: Step by Step Data Mining Guide*, SPSS.

Chapman, P., Clinton, J., Khabaza, T., Reinartz, T., and Wirth, R. (1999). *The CRISP-DM Process Model*, CRISP-DM consortium.

Ripley, B. D. (1996). *Pattern Recognotion and Neural Networks*, Cambridge University Press.

SAS Institute (2000). *Getting Started with SAS Enterprise Miner 4.1*, SAS Institute Inc.

SAS Institute (2002). *Predictive Modeling Using Logistic Regression: Course Notes*, SAS Institute Inc.

SAS Institute (2003). *Data Mining Using SAS Enterprise Miner: A Case Study Approach*, SAS Institute Inc.

[2003년 7월 접수, 2004년 2월 채택]

A Development of Cross-Sell Scoring Model

Sang-Tae Han[1] Hyuncheol Kang[2] Seong-Keon Lee[3] Yo-Cheon Jung[4]

ABSTRACT

Cross-sell models are used to predict the probability or value of a current customer buying a different product or service from the same company. Selling to current customers is one of the easiest way to increase profits and allows companies to carefully manage offers to avoid over-soliciting and possibly alienating their customers. In this study, by using the real database of an insurance company in Korea, we try to explain the steps of actual data mining process. Especially, this study aims to develop cross-sell models to predict the probability which a current customer of automobile insurance buys long-term insurance product.

Keywords: Data mining, Cross-sell, Scoring model

1) Associate Professor, Department of Informational Statistics, Hoseo University, Asan, 336-795
E-mail: sthan@office.hoseo.ac.kr
2) Senior Lecturer, Department of Informational Statistics, Hoseo University, Asan, 336-795
E-mail: hychkang@office.hoseo.ac.kr
3) Graduate Student, Department of Statistics, Korea University, Seoul, 136-701
E-mail: sklee@korea.ac.kr
4) Graduate Student, Department of Informational Statistics, Hoseo University, Asan, 336-795
E-mail: yocjung@hotmail.com