Proceedings of the Autumn Conference in 2001
and the 30th Anniversary of the Korean Statistical Society

데이터마이닝을 활용한 이탈고객 스코어링 모델 개발

한상태[1] · 이성건[2] · 강현철[3] · 유동균[4]

< 요 약 >

최근의 많은 기업에서는 방대한 고객 데이터베이스를 활용하여 자사의 경쟁력을 갖추는 방안으로써 데이터마이닝을 선택하고 있다. 본 연구에서는 데이터마이닝을 활용해 손해보험사의 데이터베이스를 분석하여 자동차보험 고객의 이탈을 방지하는 이탈고객 스코어링 모델을 개발하였다. 분석방법론으로는 의사결정나무와 로지스틱 회귀분석을 사용하였으며 기업에서의 데이터마이닝을 위한 일련의 과정을 상세히 기술하고 기업의 데이터베이스가 가지고 있는 문제점을 지적하였다.

주요용어 : 데이터베이스 마케팅, 데이터마이닝, 스코어링 모델

1. 서론

경제가 발달함에 따라 자동차의 수도 급격히 증가하고 있다. 통계청의 2001년 7월 발표에 의하면 한 가구당 자동차 보유대수는 0.88대를 넘어서고 있다고 보고된 바 있다. 자동차의 수가 늘어나면서 자동차 보험사들은 자사의 고객 유치를 위하여 보험상품을 다양화하고 타사와의 차별화 전략을 내세우고 있다. 또한 현재 자동차보험 시장은 포화상태에 있으며 신규고객 유치에서 기존고객의 관리로 마케팅 전략이 바뀌고 있는 실정이다.

이러한 변화에 발 맞추어 국내의 각 보험사들은 새로운 보험료 체계를 마련하고 고객관리를 강화하기 위해 보험가입자의 이탈 방지 및 새로운 보험상품 개발과 자동차 보험 지원 시스템을 개발하는 데이터베이스 마케팅(Database Marketing)을 통하여 경쟁력을 갖추려 하고있다(박찬욱, 2000).

본 논문의 목적은 고객 이탈 방지에 그 초점을 맞추고, 데이터마이닝을 활용하여 경쟁력있는 자동차 보험 지원 시스템과 향후 캠페인 활동의 중요한 초석이 되는 자동차 보험 이탈 고객 모형을 개발하는데 있다.

본 논문은 기업에서 진행되었던 데이터마이닝 프로젝트를 중심으로 기술하였으며 구체적인 진행순서는 CRISP-DM 방법론(Pete and Julian, 1999)에 따라 기술하였다. 데이터마이닝 소프트웨어로는 SAS사의 Enterprise Miner가 이용되었다(SAS, 1997).

2. 프로젝트 일정 및 CRISP-DM 방법론

2.1 프로젝트 일정

프로젝트는 2000년 10월에서 2001년 1월까지 4개월 동안 진행되었다. 프로젝트의 일정은

1) 호서대학교 자연과학부 수학전공 교수, (336-785) 충남 아산시 배방면 세출리 산 29-1
2) 고려대학교 대학원 통계학과 박사과정, (136-701) 서울 성북구 안암동 5가 1번지
3) 호서대학교 자연과학부 수학전공 교수, (336-785) 충남아산시 배방면 세출리 산 29-1
4) 호서대학교 자연과학부 수학과 석사과정, (336-785) 충남 아산시 배방면 세출리 산 29-1

업무파악, 분석용 마트 구축, 확증적 모형 개발, 이탈 스코어 산출 및 적용의 4단계로 구성되어 진행되었다. 전체 일정의 약 85%가 최종 분석마트의 구축에 활용되었고, 실제 모형개발 및 이탈 스코어 산출에 사용된 기간은 전체의 15%로써 분석마트의 구축 단계가 전체 프로젝트에서 매우 중요한 위치를 차지하고 있음을 알 수 있다.

2.2 CRISP-DM

본 프로젝트의 단계별 주요 과제는 CRISP-DM 방법론을 바탕으로 진행되었는데, 프로젝트의 상황 및 일정상 CRISP-DM 방법론의 일부 과제를 생략하여 진행하였다.

<표 2-1> CRISP-DM 방법론

적용 단계 구분	단계별 과제	
비즈니스 이해	1. 업무목적결정	2. 상황평가
	3. 데이터마이닝목표결정	4. 프로젝트계획수립
데이터 이해	1. 초기데이터 수집	2. 데이터기술
	3. 데이터탐색	4. 데이터품질 검증
데이터 준비	1. 데이터설정	2. 데이터선택
	3. 데이터정제	4. 데이터생성
	5. 데이터통합	
모 델 링	1. 모델링기법	2. 테스트설계
	3. 모델생성	4. 모델평가
모 델 평 가	1. 결과평가	2. 프로세스검토
	3. 향후단계결정	
전 개	1. 전개계획수립	2. 유지관리계획수립
	3. 최종보고서작성	

3. 모형정의 및 데이터 추출

3.1 목적 및 모형의 정의

자동차보험에 가입한 고객을 대상으로 보험만기가 도래했을 때 이탈하는 고객의 특성을 설명할 수 있는 자동차 보험 이탈방지 모형개발을 목적으로 하였다. 이에 따른 모형의 목표변수는 자동차 보험 만기 시점에서의 재계약 여부이며, 입력변수는 자동차 보험 계약건에 대한 만기월 기준의 고객 속성정보 및 거래정보와 과거 5년간의 거래특성을 이용하였다. 또한 자동차 보험의 성격에 따라 이탈 패턴이 다르다고 판단되어 자동차보험을 "승용", "승합용", "화물용"으로 구분하여 3개의 모형을 개발하였다.

3.2 데이터 추출

앞에서 정의한 모형에 필요한 데이터를 A보험사의 기간계 DB와 DW에서 자동차 보험 만기월 현재 기준으로 최근 5년까지의 고객 속성정보 및 거래정보와 만기월 이후 3개월 사이의 개인 고객의 갱신 정보를 추출하였다.

4. 분석용 마트 구성

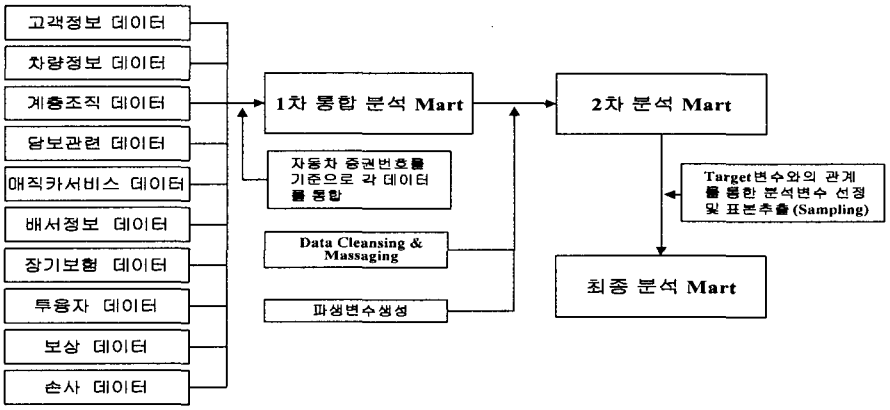
개인고객의 이탈을 예측하는데 필요하다고 판단되는 각 데이터를 A사의 기간계 시스템에서 추출한 후, 보험 증권번호를 기준으로 통합하여 1차 통합 분석 마트를 구성하였다. 데이터

정제와 파생변수를 생성하여 2차 분석 마트를 구성하였고, 마지막으로 목표변수와의 관계를 통한 분석변수 선정 및 표본 샘플링(sampling)을 통하여 최종 분석 마트를 구성하였다.

4.1 데이터 탐색

좋은 모형을 개발하기 위해서는 분석용 데이터 마트에 대한 충분한 이해가 필요하므로 빈도분석, 분할표분석, 기초통계량 분석을 통하여 데이터를 살펴보았다.

<표 4-1> 분석용 마트의 구성 흐름도



결과를 살펴보면 고객의 신상정보에 관련된 변수들이 모형개발에 매우 중요한 역할을 할 수 있음에도 불구하고 기록율이 매우 떨어져 분석변수로 사용할 수 없음을 볼 수 있다. 이는 향후 캠페인을 통한 고객정보를 획득함으로써 기록율이 향상되면 보다 안정적이고 적중률이 높은 스코어링 모형을 개발할 수 있을 것으로 기대된다.

<표 4-2> 빈도분석을 통한 결측값 확인(일부)

필 드 명	형 태	비 고
녹색면허여부	결측값 99%	분석변수에서 제외
면허취득년도	결측값 99%	
운전자경력	결측값 99%	
자녀수	결측값 99%	
학력ID	결측값 99%	
직위	결측값 99%	
선납일	99.9%가 '0'	
고객결혼상태	99.8%가 '미혼'	

4.2 데이터 정제 및 파생변수 생성

고객의 갱신에 관련된 변수들에 포함되어 있는 결측값(missing value) 및 잡음 데이터(noise data)를 파악한 후 제거 또는 적절한 값으로 대체하는 과정은 다음과 같은 단계를 거쳤다. 1단계로 결측값 및 잡음 데이터의 비율이 90% 이상이 되는 변수들을 제거하고, 2단계로는 필드간 상충, 업무적인 결측값 등은 변수간 교차 검증(cross check)를 통하여 적절한 값으로 변환하였다. 3단계는 입력오류 및 값 범위 초과 레코드들을 삭제하였다. 또한, 인수된 데이터 이외에 목표변수에 유의한 영향을 준다고 생각하는 변수를 추가적으로 생성하였는데, 이는 현업과 충분한 협의를 통해 진행되었다.

4.3 샘플링

자료의 크기가 매우 크므로 시스템 사양이 이를 뒷받침하지 못해 모형개발이 불가능하고, 표본이론(sampling theory)에 의해 전체 결과에는 영향을 미치지 않으므로 전체 100만건 중 다음과 같이 모형별로 10만건~15만건을 표본추출을 하였다.

<표 4-3> 최종 분석 마트

보 종	필드수	전체건수	재갱신건	미갱신건	갱신율	이탈율
승 용	163	150,000	88,042	61,958	58.69%	41.31%
승합용	167	90,802	41,979	48,823	46.23%	53.77%
화물용	165	150,000	74,878	75,122	49.92%	50.08%

4.4 변수선택

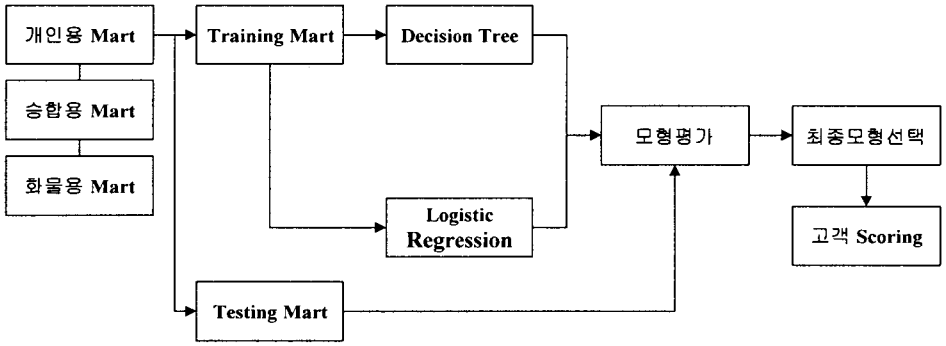
분석변수의 선정 단계에서는 목표변수와의 관계분석을 통해 각 모형별로 입력변수를 선정하였다. 목표변수를 예측하는데 연관성이 낮은 변수를 제거하고자 범주형 변수인 경우에는 카이제곱(Chi-Square)통계량을, 연속형 변수인 경우에는 t-검정을 이용하였다.

5. 이탈 스코어 모형 개발

5.1 분석흐름도

모형 개발의 분석 흐름도는 <그림 5-1>과 같이 구성하였다.

<그림 5-1> 모형 개발 흐름도

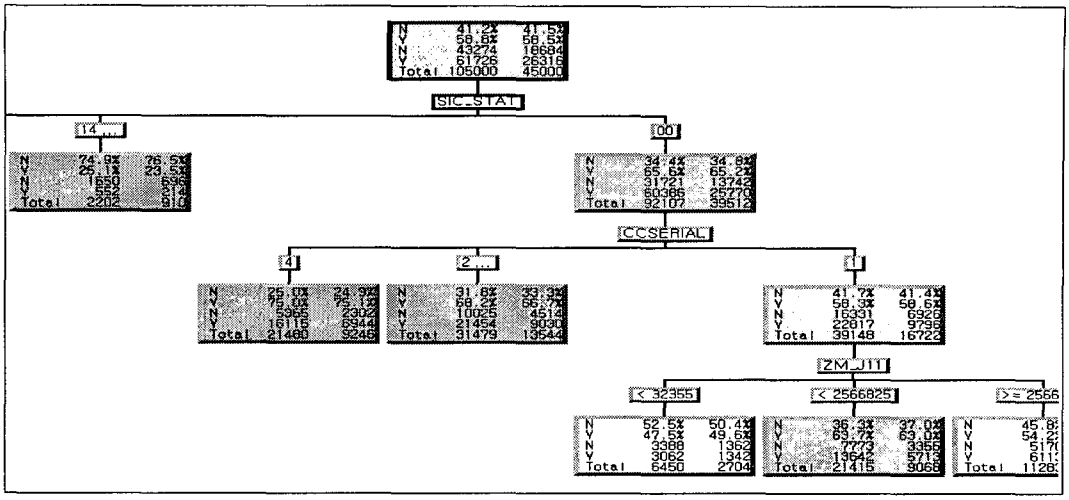


E-Miner의 데이터 파티션(data partition) 노드를 이용하여 개인용 마트, 승합용 마트, 화물용 마트의 모형개발을 위한 훈련용 부분(training data)를 70%, 모형평가를 위한 평가용 부분(test data)를 30%로 분할한 후 의사결정나무와 로지스틱 회귀모형을 개발하고 각 모형을 비교 평가하여 최적의 모형을 선택하였다(강현철·한상태 외, 2001, 최종후·한상태 외, 2001).

5.2 의사결정나무분석 결과

E-Miner의 Decision Tree 노드를 이용하였고, 알고리즘은 CHAID를 이용하였다. 승합용 마트를 분석한 결과 중 의사결정나무 결과는 다음과 같다.

<그림 5-2> 의사결정나무의 나무그림(일부)



<표 5-1> 의사결정나무에 의한 승용 모형의 오분류표

예측 / 실제	학습 결과			테스트 결과		
	미가입	가입	합계	미가입	가입	합계
미가입	15,876	27,398	43,274	6,678	12,006	18,684
가입	4,571	57,155	61,726	1,982	24,334	26,316
합계	20,447	84,553	105,000	8,660	36,340	45,000
	Prior Distribution : 58.79% 학습 정분류율 : 69.55%			Prior Distribution : 58.48% 테스트 정분류율 : 68.92%		

<표 5-1>은 학습 결과와 테스트 결과로 구분된 오분류표이다. 승용모형의 오분류표를 보면 학습 결과의 정분류율은 69.55%이고, 테스트 결과의 정분류율은 68.92%로 안정적인 모형임을 볼 수 있다.

승용 모형의 재 갱신 규칙의 일부를 살펴보면 다음과 같다.

① 계약상태가 전담보해지이면 미갱신확률이 92.7%이다.

② 계약상태가 정상이고 당사연속가입경력이 4년이면 갱신확률은 75.1%이다.

③ 계약상태가 정상이고 당사연속가입경력이 1년이고 장기수당이 32,355원 이상 2,566,825원 이하이면 갱신확률이 63.0%이다.

5.3 로지스틱 회귀분석 결과

로지스틱 회귀분석에서 변수선택법은 Stepwise 방법, 기준통계량은 AIC(Akaike's Information Criterion)를 선택하였고 이때 선택되는 변수와 제거되는 변수의 선택확률은 각각 15%, 5%로 선택하였다. 승합용 모형에 대해 로지스틱 회귀분석을 실시한 결과는 다음과 같다.

<표 5-2> 로지스틱 회귀분석에 의한 승합용 모형의 오분류표

예측 / 실제	학습 결과			테스트 결과		
	미가입	가입	합계	미가입	가입	합계
미가입	21,947	12,116	34,063	9,508	5,252	14,760
가입	8,242	21,256	29,498	3,531	8,950	12,481
합계	30,189	33,372	63,561	13,039	14,202	27,241
	Prior Distribution : 46.41% 학습 정분류율 : 67.97%			Prior Distribution : 45.82% 테스트 정분류율 : 67.76%		

5.4 모형평가 및 최종모형 선택

승용, 승합용, 화물용 모형에 대해 의사결정나무와 로지스틱 회귀분석을 실시한 학습결과와 테스트 결과의 정분류율은 다음과 같다.

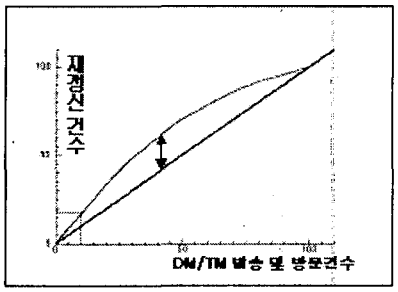
<표 5-3> 각 모형의 정분류율 비교

모형의 정분류율 비교		학습 결과	테스트 결과
의사결정나무	승 용	69.55%	68.92%
	승합용	66.80%	66.31%
	화물용	67.28%	67.18%
로지스틱 회귀분석	승 용	70.35%	69.98%
	승합용	67.97%	67.76%
	화물용	68.49%	68.30%

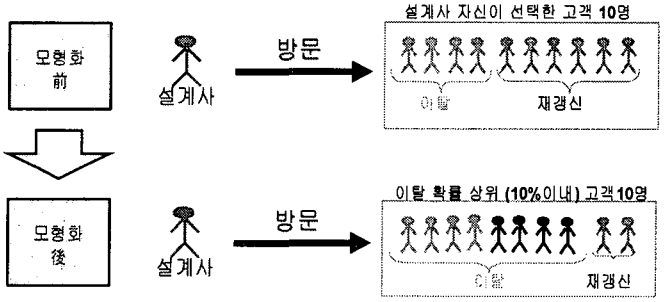
결과를 살펴보면 승용, 승합용, 화물용 모형에서 의사결정나무분석보다 로지스틱 회귀모형의 정분류율이 약간 더 우수한 것을 볼 수 있다. 하지만 뛰어난 정도가 매우 미약하고, 모형의 안정성과 향후 현장에서의 활용 용이성을 고려하여 의사결정나무분석모형을 고객 스코어링에 사용할 최종 모형으로 선택하였다.

5.5 개발된 모형을 통한 캠페인 활동

각 보종별로 개발된 모형을 토대로 하여 상대적으로 예상 갱신율이 낮은 고객을 선정하고, 그 고객을 상대로 판촉활동을 하는 캠페인을 고려하자. 향후 XX년 X월에 보험만기가 도래하는 고객을 기준으로 하는 적용데이터(applying data)를 최종 선택된 모형인 의사결정나무에 적용시켜 이탈 및 재갱신 스코어를 산출하고 산출된 고객별 스코어를 영업소 화면에 출력시켜 다음과 같이 보험설계사의 활동 효율성을 높일 수 있다.



<그림 5-3>Lift Chart



<그림 5-4>모형을 통한 캠페인 활동의 예시

6. 결론 및 토의

본 논문은 실제 데이터마이닝 프로젝트에서의 진행된 과정을 모형의 정의와 데이터의 추출에서부터 모형 개발까지 일련의 과정을 정리하였다. 또한 의사결정나무와 로지스틱 회귀분석을 이용하여 개발한 모형중 모형의 안정성과 향후 적용을 고려한 의사결정나무 모형이 선택됨을 볼 수 있었고, 데이터마이닝을 통해 얻을 수 있는 이익에 대하여 살펴보았다.

본 논문을 진행하면서 데이터마이닝 프로젝트 수행 시 중요하다고 생각한 2가지는 다음과 같다.

첫째, 프로젝트를 수행하는데 있어 현장의 마케터와 분석팀, 전산팀의 상호 이해 및 긴밀한 협조체제가 유지되어야 한다는 것이다. 즉, 현장의 요구와 감각이 모형에 반영되어야 하며, 데이터마이닝 프로젝트의 대부분의 작업이 데이터를 추출하고 조작(manipulation)하는 데에 소요되므로 전산팀의 역할 또한 매우 중요하다.

둘째, 분석 데이터의 질이다. 앞에서도 살펴보았듯이 고객의 이탈에 영향을 줄 수 있는 고객의 신상정보는 거의 획득되어지지 않고 있었고 또한 획득되어진 자료도 오류가 많아 분석에 사용할 수 없었다. 즉, GIGO(garbage in garbage out)의 사고를 가지고 프로젝트를 진행해야 할 것이다. 양질의 데이터를 신속히 제공할 수 있는 환경이 데이터마이닝 프로젝트의 성공적인 진행과 안정정인 모형을 개발하는 중요한 요소라 하겠다.

향후 자동차 보험의 차량용도와 장기보험 상품을 다각도로 고려하여 고객유형을 선정하고, 장기보험 상품과 연계된 연계 판매(cross-sell)모형 개발 등을 추후 연구 과제로 제안한다.

참고문헌

[1] 강현철 · 한상태 · 최종후 · 김은석 · 김미경 (2001). 「SAS Enterprise Miner를 이용한 데이터마이닝-방법론 및 활용-」, 서울 : 자유아카데미

[2] 최종후 · 한상태 · 강현철 · 김은석 · 김미경 · 이성건 (2001). 「SAS Enterprise Miner를 이용한 데이터마이닝-기능과 사용법」, 서울 : 자유아카데미

[3] 박찬욱 (2000). 「데이터베이스 마케팅」, 서울 : 연암사

[4] Pete Chapman and Julian Clinton (1999). Crisp-DM Process Model , *Discussion paper.*

[5] SAS Institute Inc.(1997). Data Mining Using SAS Enterprise Miner Software