데이터마이닝을 통한 고객관리데이터의 분석

(아파트고객 세분화를 중심으로)

백신정*
*고려대학교 컴퓨터 과학기술대학원
e-mail : whitebsj@hotmail.com

Ananlyzing Customer Management Data by Datamining

(Focused on Apartment Customer Classification)

Shin Jung Baek*
*Korea University Graduate School of Computer Science and Technology

요 약

기업간의 경쟁이 심화되고 정보의 중요성에 대한 인식이 확대되어 가는 상황에서 다량의 데이터로부터 가치 있는 데이터를 추출하는 CRM 데이터 마이닝은 중대한 관심사가 아닐 수 없다. 본 연구는 데이터마이닝의 여러 활용 분야 중 고객세분화를 위해 최근 많이 사용되고 있는 데이터마이닝 기법인 로지스틱 회귀분석, 의사결정나무, 신경망 알고리즘 기법들을 비교하며, 이를 실제 아파트 고객의 데이터를 이용하여 검증하고자 한다. 따라서, 아파트 고객 세분화를 위한 데이터마이닝 수행시 기법 선택의 기준과 비교 평가의 기준을 제시하는 데 연구목적 있다.

1. 서론

70 년대에는 시장전체를 대상으로 균질화된 상품과 서비스를 대량으로 공급하는 매스마케팅이 중심이었다. 매스마케팅에서는 개별고객의 선호는 중요하지 않으며 시장의 규모와 평균적 특성을 찾아내서 적합한 제품을 대량으로 생산하여 신속하게 공급하는 것이 주요관심사였다.

고객들에게 좀 더 효과적인 마케팅을 위하여 고객세분화에 의한 마케팅 기법인 세그먼트 마케팅 기법이 79 년도 중반부터 90 년 초까지 사용되었으며, 정보기술의 발전으로 고객관련 정보들과 거래실적 등을 데이터베이스로 구축하게 됨으로써 고객의 성향, 실적, 기여도 등을 분석할 수 있는 기반을 형성하게 되었다. 이러한 분석결과를 통해서 상품에 따른 고객의 분류가 가능하게 되었으며 이를 통해서 니치마켓의 발굴과 특정 고객군에게 차별적인 마케팅을 수행하는 것이 가능하게 되었다. 최근에는 고객을 가장 효과적으로 만족시킬 수 있는 퍼스널 마케팅 기법이 정보통신의 발달로 인해서 가능하게 되었다.

이러한 퍼스널 마케팅 기법 중의 하나가 바로 고객관계관리 (CRM:Customer Relationship Management)이다[8].

<그림 1.1> 마케팅의 발전과정

이렇듯 CRM 에서는 분석이라는 부분이 중요한 요소로 자리잡고 있으며, 그중 가장 대표적인 분석 기법이 데이터마이닝이다.

따라서 본 연구의 목적은 첫째 데이터마이닝에 관한 이론적 측면을 각종 문헌을 통하여 정리하고, 둘째 A 사의 CRM 데이터를 사례로 아파트 업계 고객의 체계적인 분류 기준을 제시하고 분석기법을 비교 평

가하여, 향후 아파트 업계의 고객분석, 관리하는데 도움이 되고자한다.

본 연구는 4 개의 장으로 구성되어 있으며, 각 장의 내용은 다음과 같다. 1 장은 서론부분, 2 장은 관련된 연구를 통해 CRM 과 데이터마이닝의 이론의 소개로서, 정의, 활용분야, 널리 적용 되는 기법들을 기술한다. 3 장은 실제데이터를 이용한 세분화 기법을 적용하여 보고 결과를 기술하며, 전략을 도출한다. 4 장은 결론을 맺는다.

2. CRM 을 위한 데이터마이닝

2.1 데이터마이닝의 정의

데이터마이닝 기법이란 대량의 데이터로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 과정을 말한다. 여기서 정보는 묵시적이고 잘 알려져 있지 않지만 잠재적으로 활용가치가 있는 정보를 말한다. 다시 말해 데이터마이닝이란 기업이 보유하고 있는 일일 거래자료, 고객자료, 상품자료, 마케팅활동의 피드백 자료와 기타 외부자료를 포함하여 사용 가능한 데이터를 기반으로 숨겨진 지식, 기대하지 못했던 패턴, 새로운 법칙과 관계를 발견하고 이를 실제 경영의 의사결정 등을 위한 정보로 활용하고자 하는 것이다[3].

2.2 데이터마이닝 주요기법

데이터마이닝은 알고자 하는 정보에 따라 작업 유형이 결정된다. 작업유형은 크게 연관규칙(associate-on) 연속규칙(sequence), 분류규칙(classification), 데이터 군집하(clustering) 등 4 가지 유형으로 나누어진다. 그리고 이 네 가지 작업유형을 지원하는 데이터마이닝 기법은 전통적인 통계기법 (예: 회귀분석, 판별분석), 의사결정나무, 신경망, 동시발생 매트릭스(Co-Occurrence Matrix), K-평균 군집화(K-Means Clustering)기법 등이 있다[3].

이 중에서도 의사결정나무와 신경망 기법등과 같이 인공지능에 기반을 둔 기법들이 대표적인 데이터마이닝 기법이라고 보고 있는 학자들이 있다.

위에서 열거한 여러 가지 기법들 가운데 본 연구에서 적용한 의사결정나무, 신경망, 회귀분석에 대해 살펴보겠다.

의사결정나무는 데이터마이닝의 분류 작업에 주로 사용되는 기법으로, 과거에 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 패턴, 즉 분류별 특성을 속성의 조합으로 나타내는 분류모형은 새로운 레코드를 분류하고 해당 부류의 값을 예측하는데 사용된다.

인공신경망은 인간이나 동물들이 가지고 있는 생물학적인 뇌 신경세포를 모델화하여 인공적으로 지능을 만드는 것이다. 즉 인간의 뇌에 존재하는 생물학적 신경세포와 이들의 연결 관계를 단순화시켜 수학적으로 모델링하여 인간의 두뇌가 나타내는 지능적 형태를 구현하는 것으로 마디와 고리로 구성된 망구조를 모형화하고, 의사결정나무와 마찬가지로 과거에 수집된 데이터로부터 반복적인 학습과정을 거쳐 데이터에 내재되어 있는 패턴을 찾아내는 모델링 기법이다[7].

회귀분석은 한 변수(목표변수)가 다른 변수(입력변수)들에 의해서 어떻게 설명(explanation)또는 예측(prediction)되는 지를 알아보기 위해 자료를 적절한 함수식으로 표현하여 분석하는 통계적인 방법이다[1]. 회귀분석은 여러 가지 분류로 나누는데 그 중 선형, 비선형으로 나뉠 수 있다. 본 연구는 비선형회귀분석의 대표적인 기법인 로지스틱 회귀분석으로 분석한다.

2.3 기법 비교

데이터마이닝을 분석목적에 따라 크게 기술적 마이닝(Descriptive Mining)과 예측적 마이닝(Predictive Mining)으로 분류한다. 기술적 마이닝은 주어진 데이터를 설명 하는 패턴을 찾아내는 것이 주목적이며 찾아낸 패턴을 사용자의 이해를 위 해 표현, 설명하는 작업으로써 연관규칙(Association Rules), 순차적 패턴 (Sequential Patterns), 군집분석(Clustering)등이 이에 해당되며 연관규칙과 순차적 패턴을 묶어 동시발생 매트릭스(Co-occurrence Matrix) 기법이라고 도 한다. 이러한 기술적 마이닝은 분석대상 데이터에 목표변수가 존재하지 않으므로 자율학습 마이닝(Unsupervised Mining)이라고도 한다. 예측적 마이닝은 주어진 데이터에 근거하여 모형을 만들고 이 모형을 이용하여 새로 운 경우에 대한 예측을 하는 작업으로 분류(Classification) 및 예측(Prediction)등이 이에 해당된다. 이러한 예측적 마이닝은 분석대상 데이터에 목표변수와 예측변수가 함께 존재하므로 지도학습 마이닝(Supervised Mining)이라고도 한다. 군집분석에 적용되는 알고리즘에 가장 많이 사용되 는 것으로 전통적 군집방법의 하나인 K-means 군집분석 알고리즘과 신경망 분석의 하나인 Kohonen Network 알고리즘이 있으며, 분류 및 예측에 가 장 많이 사용되는 것으로는 의사결정나무와 신경망 기법이 있다.

아래 <표 2.1> 는 데이터 마이닝의 기술적, 예측적 마이닝 기법을 구분하여 도표로 정리한 것이다[6].

<표 2.1> 데이터마이닝 기법 비교

기술적 마이닝 (Descriptive Mining)		예측적 마이닝 (Predictive Mining)
동시발생 매트릭스	군집분석	분류 및 예측
연관규칙 순차적 패턴	전통적 군집분석 신경망	의사결정나무 신경망

3. 사례연구 : 아파트 고객 세분화

3.1 문제정의

현재 아파트 업계는 IMF 의 마감과 금리하락이란 원인등 으로 2001 년 경 부터 오르기 시작한 아파트 가격은 많은 전문가의 가격하락 예상을 뒤 엎고 계속해서 치솟아왔으며, 유례없는 아파트계의 호황을 지난 몇 년간 누려왔다고 해도 과언이 아니다. 하지만

강남의 재건축 아파트 가격의 상승으로 인하여 정부에서는 각종 정책 내 놓고 있고, 원자재 가격상승, 분양가를 공개하라는 움직임 등이 최근 몇 년의 아파트 시장을 뒤돌아 볼때 가장 불안한 상황이 아닐 수 없다.

하지만 최근 몇년간 꾸준한 가격상승과 아파트 경기의 상승은 기존의 건설회사라는 딱딱한 이미지를 벗어던지고, 건강아파트 도입, 아파트 브랜드화, 거주의 개념만이 아닌 삶의 질을 높여주는 아파트 이미지 등 한차원 높은 마케팅 개념이 도입되도록 부추겨 주고 있는 실정이기도 하다. 이러한 마케팅의 많은 방법론 중에 고객을 잘 관리하고 효과적인 마케팅 접근방법 등을 도출해내는 CRM 은 아파트 업계에서도 반드시 필요한 사항이라 여겨진다.

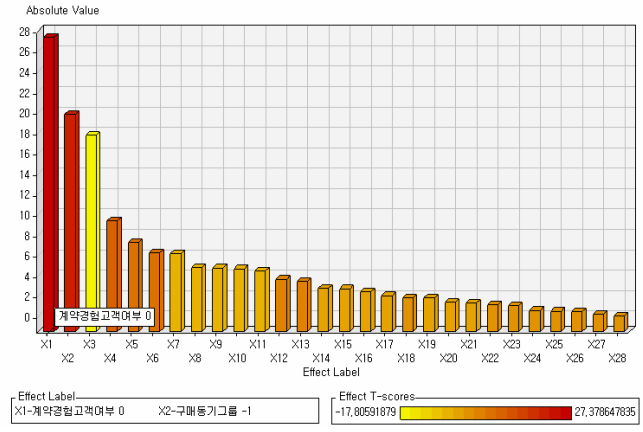
다른 나라와 달리 우리나라의 대표적인 주택인 아파트는 분양에서 입주까지 3~5 년의 장기간이 걸리고 한번 구매한 고객이 반복하여 신규아파트를 구매하기까지에는 평균적으로 10 년 이상의 장기간이 소요되며, 타 업계와 달리 업계의 특성상 개인고객의 개념보다는 가족고객의 개념이라는 특수성이 존재한다. 또한 거의 모든 고객들은 2-30 대에 내집마련의 꿈을 꾸며, 40-50 대의 고객들은 고교평준화로 인해 자녀들의 학군을 고려하고 있으며, 고령화 사회에 달려가고 있는 현실은 실버타운등을 비롯한 주택의 새로운 대안을 내 놓아야만 하는 현실이기도하다. 이것은 거의 모든 국민이 아파트에 관심을 가지고 있다는 말이 되기도 한다. 이러한 이유 때문에 아파트 업계의 CRM 전략은 기존의 통신산업이나 금융업계와는 전혀 다른 접근이 필요 하지만, 아직까지 아파트 업계는 이제 막 CRM 을 도입하거나, 도입 하고 있는 걸음마를 시작하고 있는 현실이기도 하다

이러한 이유로 본 연구에서는 아파트 고객세분화를 데이터마이닝을 이용하여 A 사의 아파트 고객 데이타에 실제로 적용하여 다각적으로 분석해 보고자한다.

3.2 실험구성

본 연구에서는 A 사의 CRM 데이터 총 30 만개 중 Sample data 6 만개 정도를 사용하여 연구를 실시하였으며, 데이터마이닝 통계 패키지 툴인 SAS Enterprise Miner 로 실험하였다.

<그림 3.1> 아파트 고객 데이터



데이터는 학습용으로 40%, 평가용으로 30%, 검증용으로 30%로 분할하여 사용하였고, 86 개의 변수 중 중요한 22 개의 변수를 추려서 실험하였다.

변수추출은 첫 번째로는 경험에 의한 주관적인 판단을 전제로 하였으며, 추출한 변수를 바탕으로 검증은 T-점수(T-score)를 그래프로 표현하여 회귀계수가 양의 값을 가지는 변수를 <그림 3.1>와 같이 선정하였다.

또한 목표(Target) 변수로는 최근 2 년내 아파트 계약여부를 선택하였으며, 4 년 전의 2 년간의 데이터를 입력변수로 선택하여 분석을 실시하였다. 데이터와 형식은 아래 <표 3.1>과 같다.

<표 3.1> 아파트 고객 변수

번호	이름	타입	변수설명
1	TARGET 변수 (향후 2 년내 계약여부)	binary	1:계약, 0:비계약
2	2 회이상계약여부	binary	
3	거주지역구	nominal	
4	거주지역구분	nominal	
5	계약경험고객여부	binary	
6	구매동기그룹	ordinal	
7	분양희망여부	binary	
8	분양희망지역그룹	nominal	
9	성별	binary	
10	시공권조합원여부	binary	
11	연령	interval	
12	입주경험고객여부	binary	
13	재계약여부	binary	
14	전매경험고객여부	binary	
15	청약통장유무	binary	
16	총계약횟수	ordinal	
17	최근 1 년이내계약여부	binary	
18	최근 1 년이내입주여부	binary	
19	최근 1 년이내전매여부	binary	
20	최근 2 년이내계약여부	binary	
21	최근 2 년이내입주여부	binary	
22	최근 2 년이내전매여부	binary	

이 데이타의 목표변수의 비율은 계약 92.85%,비계약 7.15%이다.

3.3 실험결과

각 모형의 정확성은 오분류표를 이용하여 평가하였으며, 오분류표는 목표변수의 실제 범주와 모형에 의해 예측된 분류 범주 사이의 관계를 나타내는 표로써 대각선 원소에는 정분류를 비대각 원소에는 오분류를 나타내어 목표변수의 범주별로 이를 제대로 분류한 빈도와 그렇지 못한 빈도를 함께 제시한다. 아래의 표에서 정분류 빈도는 오분류표에서의 대각선 원소의 합으로 목표변수의 범주를 제대로 분류한 빈도를 나타내며 오분류 빈도는 비대각 원소의 합으로 잘못 분류한 빈도를 나타낸다. 정확도(정분류율)는 전체 데이터 셋에서 정분류 빈도가 차지하는 비율을 의미하

며 오분류율은 전체 데이터 셋에서 오분류 빈도가 차지하는 비율을 의미한다[6].

3.3.1 로지스틱 회귀분석
검증결과 :

<표 3.2> 로지스틱 회귀분석 모형 오분류표

	예측결과			
실제반응		비계약	계약	합계
	비계약	21809	67	21876
	계약	1835	123	1958
	합계	23644	190	23834

- 정분류율 : 92.02%
- 오분류율 : 7.98%

로지스틱 회귀분석 모형의 정확도를 측정한 결과 정분류율은 92.02%, 오분류율은 7.98%로 나타났다.

3.3.2 의사결정나무(Decision Tree)
검증결과 :

<표 3.3> 의사결정나무 모형 오분류표

	예측결과			
실제반응		비계약	계약	합계
	비계약	21789	87	21876
	계약	1807	151	1958
	합계	23596	238	23834

- 정분류율 : 92.05%
- 오분류률 : 7.95%

의사결정나무 모형의 정확도를 측정한 결과 정분류율은 92.05%, 오분류율은 7.95%로 나타났다.

3.4 결과분석
아파트고객 세분화를 위한 모형에 대하여 오분류율을 비교해 보고 평가해 보았다.
아래 <표 3.5>는 로지스틱 회귀분석과 의사결정나무의 정분류율, 오분류율을 구한 것이며, 내용을 보면 오분류율 평가 기법을 기준으로 의사결정나무 모형이 가장 우수한 기법이라고 할 수 있다.

<표 3.4> 각 모형 정분류율, 오분류율 비교표

	로지스틱 회귀분석	의사결정나무
정분류율	92.02	92.05
오분류율	7.98	7.95

4. 결론

본 연구는 아파트 고객분류의 기법에 대해 데이터마이닝에 있어 널리 활용되는 로지스틱 회귀분석, 의사결정나무을 사용하여 고객세분화에 영향을 미치는 요인과 분류하는 기법에 대하여 연구하였다. 그리고 모형의 특성을 검증하기 위해 아파트 회사 고객 데이터를 이용하여 고객 세분화 모형을 구축하고 그 결과를 비교하였다. 각 기법의 평가결과 거의 차이는 없었지만 의사결정 나무 모형이 다른 모형에 비해 우수함을 보였고, 고객을 세분화하는데 가장 중요한 변수로는 계약경험고객여부와 구매동기그룹으로 나타났다.
이러한 특징은 전문가라 할지라도 수동으로 쉽게 찾아내기 어려운 특징이다. 본 연구는 데이터마이닝에서 아파트 고객세분화의 개념을 정립하였으며 고객세분화를 위한 데이터마이닝 기법 및 알고리즘의 비교 기준을 제시했다는데 그 의미가 있다. 또한 본 연구를 통해 아파트 고객세분화에 대한 모형을 이용하여 아파트계약 잠재고객을 찾아내는 등의 마케팅 전략을 도출할 수 있을 것이다.
그러나 고객세분화 중 데이터 특성에 대한 실증 연구가 이루어지지 않았고 또한 하나의 데이터에 대한 실증 비교 라는게 문제가 있다. 따라서 향후에는 세분화에 대한 모형 평가 방법의 제시 및 검증과 모의 자료를 이용한 데이터 특성에 관한 실증 연구 그리고 다양하고 충분한 데이터 셋을 이용한 실질적 의미의 검증이 필요하다.

참고문헌

[1]강현철.한상태.최종후.김은석.김미경,“SAS Enterprise Miner 4.0 을 이용한 데이터마이닝 - 방법론 및 활용”, 자유아카데미, 2001.
[2]강현철.한상태.최종후.김은석.김미경,이성건 “SAS Enterprise Miner 4.0 을 이용한 데이터마이닝 - 기능 및 사용법”, 자유아카데미, 2001.
[3]장형진.최성.한정란.이기민, “데이터마이닝을 이용한 eCRM", 정보처리학회보, 제 8 권 제 6 호(2001.11)
[4]이도현, “데이터마이닝을 이용한 CRM", 정보과학회지, 제 18 권 제 11 호(2000.11)
[5]강한구, "이탈 고객 분류를 위한 데이타마이닝 방법의 비교 연구", 동의대학교 전산통계학과 전산과학전공, 2002.12
[6]조혜정, "고객세분화를 위한 데이터마이닝 기법 비교." 동아대학교대학원 경영정보학과, 2001.12
[7]도기윤, "데이터마이닝 기법을 이용한 은행계 신용카드 연체고객 분석", 고려대 경영대학원, 2002
[8]박주석, “정보기술과 마케팅의 변화 : CRM ” 대한산업공학회 2000
[9]이동통신고객 분류를 위한 의사결정나무(C4.5)와 신경망 결합 알고리즘에 관한 연구. 고려대학교 산업시스템 정보공학과 이극노 [2003. 6]
[10]김재찬, “다중판별분석을 이용한 주택산업의 마케팅 전략에 관한 연구”, 한양대학교 대학원 건축공학과, 1998
[11]이경미, “컨조인트 부분효용 함수를 이용한 시장세분화에 관한 연구”, 한양대학교 대학원 경영학과, 1999