사례를 기반으로 한 신문 산업에서의 고객 이탈 예측 모형 구축

양 승 정* · 이 종 태**
*동국대학교 산업기술연구원 · **동국대학교 산업시스템공학과

Development of churn prediction model in a newspaper based on real case

Seung Jeong Yang* · Jong Tae Rhee**
*Research Institute for Industrial Technology, Dongguk University
**Department of Industrial System Engineering, Dongguk University

Abstract

What is CRM(Customer Relationship Management) means that planning, executing, and re-accessing the marketing strategy based on the customer character by analyzing the material related to customers. That is CRM is a strategy of customer service on the base of data.

In the case of the telecommunications and a newspaper, there are restricted application of CRM, because they are provided services by paying a given amount of money within a given period of time.

This paper develops CRM model(churn prediction model) that can apply to a newspaper.

For model-building, real data were used which were collected from one of the major a newspaper company in Korea. Also, this paper verifies the efficient result.

Keywords: Customer Relationship Management, Data Mining, Churn Prediction Model

1. 서 론

고객관계관리(Customer Relationship Management, CRM)는 고객에 대한 정확한 이해를 바탕으로 고객이 원하는 제품과 서비스를 지속적으로 유지시키고 결과적으로 고객 평생가치를 극대화하여 수익성을 높일 수 있는 통합된 고객관계 관리 프로세스를 효과적으로 운영하는 것이다.[1, 9]

이는 단순히 시장점유율 확보를 위한 고객유도 활동을 넘어 양질의 서비스와 경험을 제공함으로써 반복구매와 충성고객을 유지하여 이른바 고객점유율을 향상시키기 위한 기업의 장기적이고 전략적인 노력이라 할 수 있다.[3, 5]

CRM이 가장 많이 이용되고 있는 산업분야는 신용카드 회사, 은행, 온라인 쇼핑몰, 보험사 등 다양한 컨텐츠 구매 기회를 제공하는 산업형태를 가지고 있어 새로운 영업 컨텐츠가 계속적으로 제시되며, 기존 고객들로 하여금 반복 구매나 교차구매를 유도할 수 있는 산업분야들이다.[4, 6]

이에 비해 일정 기간, 일정한 사용료를 지불하고 컨텐츠를 사용하는 통신 산업이나 신문 산업의 경우는 반복 구매나 고객의 충성도를 향상시키기 위한 고객관계관리 시스템을 적용하는데 많은 제약이 따른다.[7]

이는 특히 신문 산업의 경우 고객에게 제공할 수 있는 구매기회는 '신문'이라는 단일 상품뿐이기에 더욱 그러하다. 그러므로 신문 산업이 시장경쟁에서 직면하게 되는 여러 가지 한계와 위험을 극복하기 위한 안전경영의 한 부분으로 CRM 시스템 도입이 중요하게 대두된다.

2007년 5월 접수; 2007년 6월 수정본 접수; 2007년 6월 게재확정

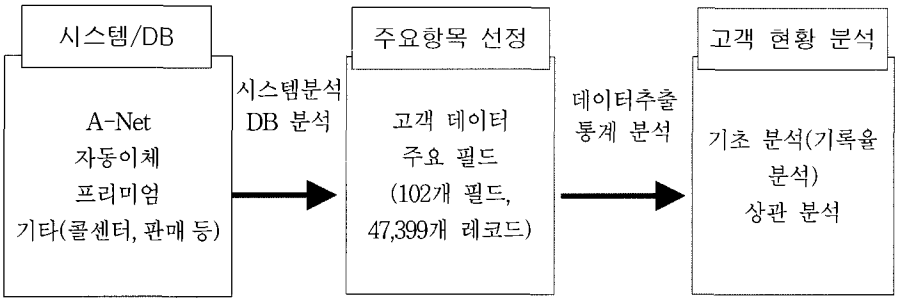
이에 본 논문에서는 국내 3대 신문사 중 하나인 A사의 실제 고객 데이터를 이용하여 신문 산업에 맞는 CRM을 실시한 뒤, 그 결과를 분석해 보고자 한다. 최종적으로 A사의 신문 구독을 중지하는 고객들을 예측하는 구독중지 예측 모형을 구축한 뒤 실제 구독중지율과 비교하여 구축 모형을 검증한다.

2. 분석대상 및 절차

A사 고객 데이터 분석을 위해 A사에서 보유하고 있는 기존 시스템인 A-Net, 자동이체, 프리미엄, 기타 시스템에 대해 시스템 분석 및 DB 분석을 통해 주요 분석 필드 및 신문 구독 중지에 미치는 영향력 항목을 선정하였다. 그리고 선정된 항목에 대해 2003년 1월부터 2005년 5월 총 29개월 동안 A사 신문을 구독하고 있던 독자들 중 일산 지역 거주자 정보, 총 47,399개의 데이터 레코드를 추출하였다.

추출된 고객 데이터를 토대로 기초 분석, 상관 분석을 수행하여 이탈 예측 모형 구축을 위한 중요 변수를 도출하였다.

<그림 1>은 이러한 데이터 분석 절차를 나타내고 있다.

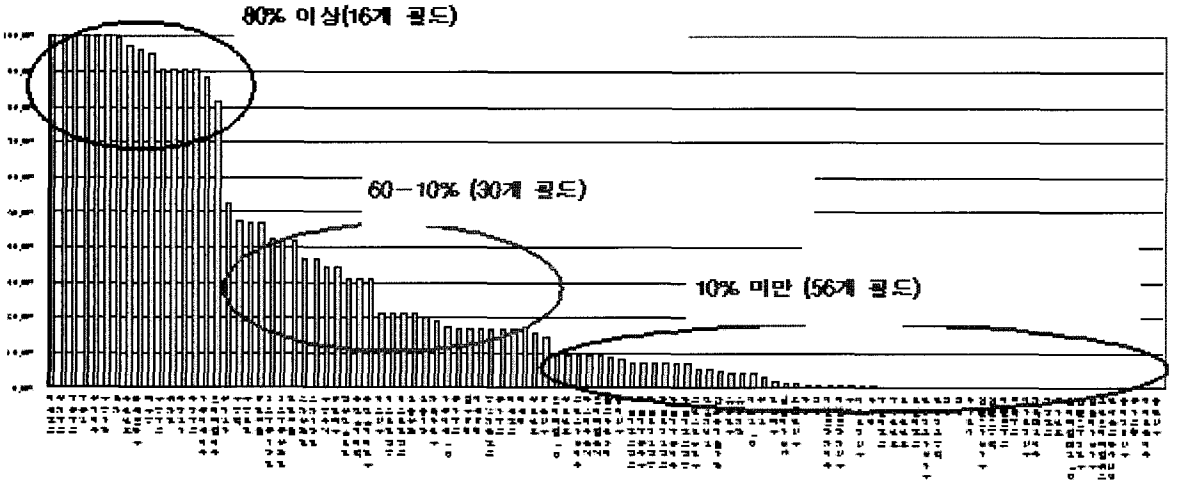


<그림 1> 원데이터 분석 절차

2.1 데이터 기록 현황

A신문사의 기존 시스템인 A-Net, 자동이체, 프리미엄, 기타 시스템 DB의 고객 데이터의 값이 데이터베이스에 존재하는 비율을 살펴보았다. 분석 대상 데이터의 총 레코드 개수는 47,399개이며, 필드수는 102개이다.

총 102개의 필드들에 대한 데이터 기록율을 살펴보면 <그림 2>와 같다. 80% 이상의 기록율을 갖는 필드는 16개, 60%에서 10%의 기록율을 갖는 필드는 30개, 10% 미만의 기록율을 갖는 필드는 56개이다.



<그림 2> 고객데이터 기록율

2.2 빈도 분석

총 102개의 변수 중 의미 있는 주요 변수를 도출하기 위해 빈도분석을 실시하였다. 각 변수별로 목적 변수인 '구독 중지 여부'의 빈도와 기록율을 비교하였다. 직접적으로 구독 중지 여부에 영향을 미치지 않는 변수는 탈락시켰으며, 기록율이 10%대로 낮은 변수 또한 탈락시켰다.

<표 1>은 신문대금의 자동이체 신청 여부와 구독 중지 여부의 관계를 나타내고 있다.

<표 1> 자동이체 여부에 따른 구독 중지 빈도 분석 결과

값구분	Y	N	NULL	합계
빈도	6,752	35,149	5,498	47,399
비율	14.25%	74.16%	11.60%	100.00%

신문대금 지불 방법을 자동이체로 신청해 놓은 고객의 경우 신문 구독을 중지한 경우는 총 47,399명 중 6,752명으로 14.25%를 차지했고, 신문 구독을 중지하지 않은 경우는 35,149명으로 75.16%나 되었다.

다시 말해 신문대금 지불 방법을 자동이체로 설정해 놓는다면 이탈할 확률이 그렇지 않은 경우보다 훨씬 적어진다는 것을 알 수 있으며, '자동이체 여부'는 목적변수인 '구독 중지'에 대단히 큰 영향을 미치고 있다는 것을 알 수 있다.

총 102개의 변수에 대해 위와 같은 빈도분석을 모두 실시하여 목적변수인 구독중지에 영향을 미치는 변수를 추출하였다.

빈도분석과 기록율을 감안하여 최종적으로 분석에 사용하게 될 19개의 변수를 추출하였으며, 이에 대한 일반적인 현황은 <표 2>와 같다.

<표 2> 분석에 사용하게 될 주요 변수

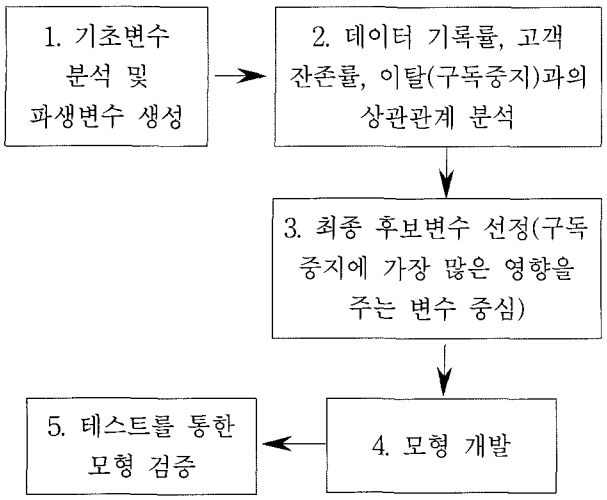
	레이블	타입	measure	코드값 및 데이터 범위
1	독자번호	String	Nominal	ID Key
2	구독일	Date	Scale	2004-01-08 (Date형)
3	구독기간	Date	Scale	04-01 (Date형)
4	고객등급	String	Nominal	E, G, P, H1 .....
5	이사신청 횟수	Numeric	Nominal	1
6	주거구분	String	Nominal	01, 02, 03, 04, 05, 06
7	프리미엄여부	String	Nominal	Y, N
8	학습지 구독 여부	String	Nominal	Y
9	학습지 구독 부수	Numeric	Numeric	1, 2, 3
10	결혼 년차	String	Nominal	1부터 72
11	자동이체 여부	String	Nominal	Y, N
12	성별	String	Nominal	남, 여
13	연령대	String	Scale	10대, 20대, 30대 ....
14	유가월	String	Nominal	200501 .....
15	의무구독기간	Date	Nominal	0부터 21
16	휴독기간	String	Nominal	0부터 1637
17	중지여부	String	Nominal	Y, N
18	아파트 평수	String	Nominal	9부터 130
19	연체 금액(미수금)	Numeric	Numeric	1000, 3000, ... , 360,000

3. 고객이탈 모형(구독중지예측 모형) 개요

3.1 고객이탈 예측 모형 구축

3.1.1 고객이탈 예측 모형 프로세스

고객이탈 예측 모형 프로세스는 변수 분석을 통해서 이탈모형의 입력변수를 선택하고, 모형을 개발하고, 평가하는 작업을 말하며, <그림 3>에서 자세히 나타내고 있다.



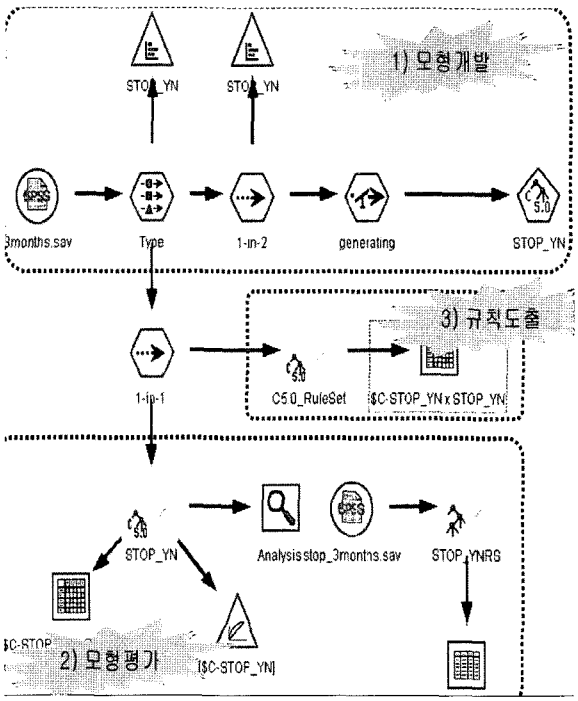
<그림 3> 이탈 모형 프로세스

3.1.2 고객이탈 예측 모형 설계

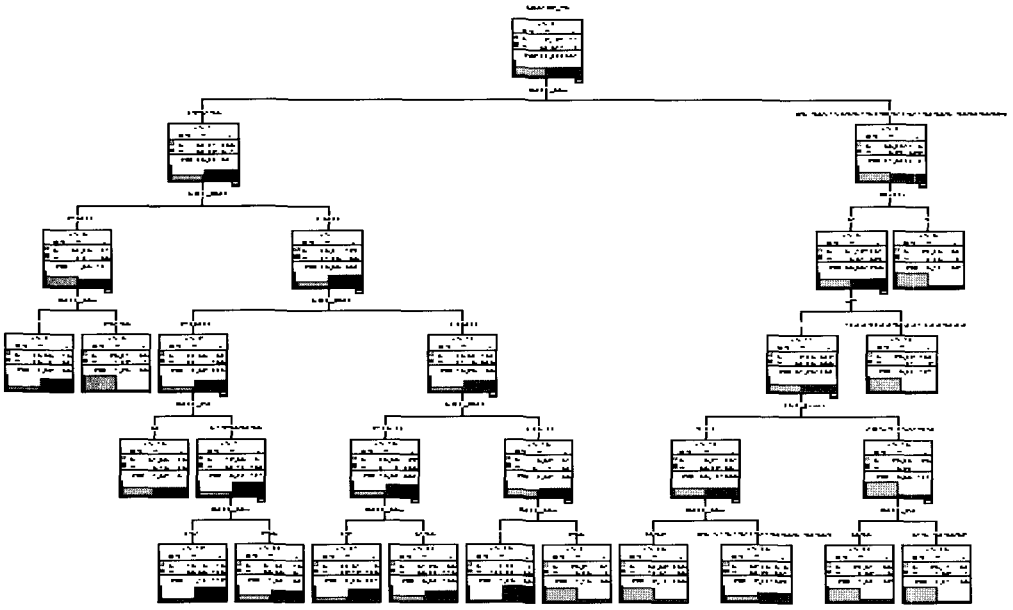
구독 중지 예측 모형의 입력변수는 통계분석을 통해 중지에 가장 영향을 크게 주는 변수를 도출하여 선정하였다. 사용된 고객 데이터는 중지예측의 정확도를 위해 1개월마다 중지율을 비교하였고, 매달 기준으로 고객들을 추출하였다. 예를 들어 2005년 4월 고객의 중지율 예측을 위해 2003년 1월부터 2005년 3월까지의 정상고객을 대상으로 모형을 설계하였다.

구독 중지 예측 모형 개발 프로세스는 모형개발 프로세스, 모형평가 프로세스, 이탈규칙 도출과 같이 3가지로 구성되며, 이를 그림으로 나타낸 것이 <그림 4>이다. 이때 사용된 데이터마이닝 툴은 클레멘타인 8.0이며, 의사결정나무분석 중 하나인 C5.0분석 방법을 이용하였다.[10, 11]

<그림 5>는 C5.0을 이용하여 도출된 구독 중지 모형을 나타내고 있다.



<그림 4> 구독 중지 예측 모형 개발 프로세스



<그림 5> 의사결정나무(C5.0)를 이용한 구독 중지 모형

3.2 고객이탈 예측 모형 검증

3.2.1 정오분류표를 이용한 검증

구독 중지 모형의 검증을 위하여 정오분류표를 이용하였다.

정오분류율(hit ratio)은 입력 변수의 값을 갖고 각 사례의 실제 결과를 예측하는 분류적 예측 모형의 성능 평가에 가장 일반적으로 사용되어온 방법이다.[2]
우선, 2004년 1월 경우 정오분류표는 <그림 6>과 같다.

<그림 6>을 보면 구독 중지 예측 수는 총 4211명이고, 그 중 실제 구독을 중지한 고객은 591명으로 14.035%를 차지하고 있다.

2004년 1월의 실제 구독 중지율은 4.72%이며, 이에 비하면 예측율이 약 3배 높은 것을 알 수 있다.
또한 계속 신문을 구독할 고객의 예측 정확도는 99.84%이다.

$C-STOP_YN		N	Y	Total
N	Count	8586	13	8599
	Row %	99.849	0.151	100
Y	Count	3620	591	4211
	Row %	85.965	14.035	100
Total	Count	12206	604	12810
	Row %	95.285	4.715	100

<그림 7> 정오분류표 결과(2004년 1월)

3.2.2 이득도표를 이용한 검증

좀 더 자세한 결과의 검증을 위하여 이득도표를 살펴보았다.

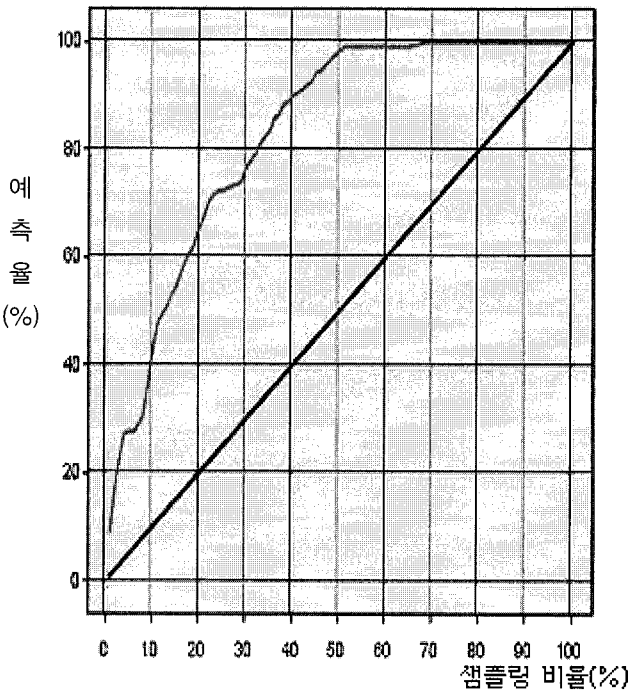
이득율(gain ratio)이란 마케팅 프로모션이나 이벤트 시행 시 예산이나 기타 문제로 모든 고객을 대상으로 하기 어려울 경우, 마케팅의 효과가 높을 것으로 예상되는 고객군을 얼마나 정확하게 예측하는 가를 평가하는 기준이다.[2]

<그림 7>은 2004년 1월의 이탈 예측 모형에 대한 이득도표를 나타내고 있다.

이득도표를 통해 고객 이탈 예측 모형을 평가하면, 50%에 해당하는 이득율은 99% 정도로 예측되고 있다.

이것은 이탈 예측 모형이 50%의 중지고객을 분류하면 전체 중지 고객의 99%를 정확하게 예측할 수 있다는 평가 척도가 되는 것이다.

샘플링 고객 수가 40% 이상이 되면 중지 고객을 90% 이상 예측할 수 있는 것이다.



<그림 8> 이득도표 결과

3.2.3 실제 중지결과를 이용한 검증

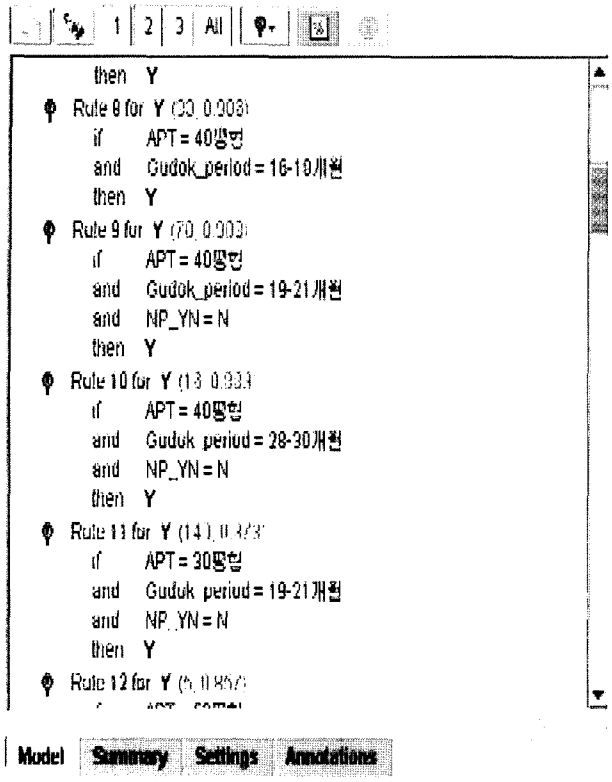
실제 신문 구독을 중지한 결과를 이용하여 고객 이탈 예측 모형의 정확도를 살펴보고자 한다.

우선 중지예측 모형을 이용해 각 기간별 구독 중지 예측 인원을 파악한다. 이때 정오분류표와 도출된 규칙을 이용한다.

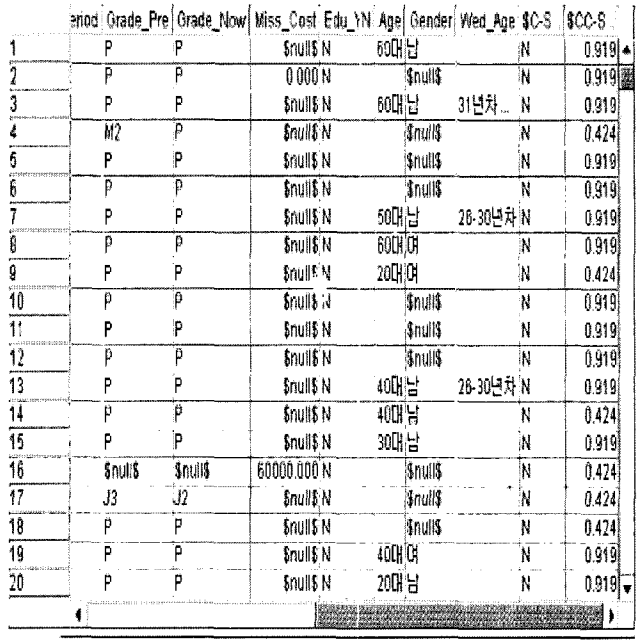
그리고 구독 중지 가능성이 높은 순서대로 정렬한 뒤 실제 중지한 인원과 비교, 분석한다.

<그림 8>과 같이 구독중지 예측 모형에 의해 규칙을 도출한 뒤 그 규칙에 따라 구독중지 예측 인원을 파악한다. 또한 중지가능성이 높은 순서대로 나열한 뒤 실제 신문 구독을 중지한 고객과 비교하여 예측율의 정확도를 검증한다.

<그림 9>는 실제 중지한 고객 수와 중지예측율이 높은 고객의 수를 비교하기 위한 리스트를 나타내고 있다.



<그림 8> 중지예측 모형에 의해 도출된 규칙 집합



	eriod	Grade_Pre	Grade_Now	Miss_Cost	Edu_YN	Age	Gender	Wed_Age	$C-S	$CC-S
1		P	P	$null$	N	60대	남		N	0.919
2		P	P	0.000	N		$null$		N	0.919
3		P	P	$null$	N	60대	남	31년차 ...	N	0.919
4		M2	P	$null$	N		$null$		N	0.424
5		P	P	$null$	N		$null$		N	0.919
6		P	P	$null$	N		$null$		N	0.919
7		P	P	$null$	N	50대	남	26-30년차	N	0.919
8		P	P	$null$	N	60대	여		N	0.919
9		P	P	$null$	N	20대	여		N	0.424
10		P	P	$null$	N		$null$		N	0.919
11		P	P	$null$	N		$null$		N	0.919
12		P	P	$null$	N		$null$		N	0.919
13		P	P	$null$	N	40대	남	26-30년차	N	0.919
14		P	P	$null$	N	40대	남		N	0.424
15		P	P	$null$	N	30대	남		N	0.919
16		$null$	$null$	60000.000	N		$null$		N	0.424
17		J3	J2	$null$	N		$null$		N	0.424
18		P	P	$null$	N		$null$		N	0.424
19		P	P	$null$	N	40대	여		N	0.919
20		P	P	$null$	N	20대	남		N	0.919

<그림 9> 도출된 규칙을 이용하여 실제 중지한 인원과 비교하기 위한 리스트

위와 같은 과정을 거쳐 이탈 예측 모형을 검증한 결과를 <표 3>에 나타내고 있다. 2003년 6, 9, 11월, 2004년 1, 3, 6, 11월, 2005년 1, 4월의 각 중지율과 신문 구독 중지 예측율, 예측에 의한 정확율, 중지율에 대한 정확율을 나타내고 있다.

<표 3> 중지 결과를 이용한 예측의 검증 비교

	2003/6	2003/9	2003/11	2004/1	2004/3	2004/6	2004/11	2005/1	2005/4
중지율	8.85%	8.80%	7.94%	4.72%	10.99%	5.14%	5.46%	4.15%	3.45%
예측(Y-Y)	72.89%	45.84%	26.27%	19.51%	34.19%	18.09%	18.68%	9.75%	6.91%
예측에 의한 정확률	69.94%	60.%	52.28%	41%	57%	54%	53.15%	33%	44.43%
중지율에 의한 정확률	78.47%	77.23%	76.04%	75%	88%	61.81%	89.70%	82.27%	85.44%

<표 3>의 결과를 살펴보면, 2003년 1월부터 2005년 5월까지 평균 구독 중지율은 6.61%이며, 고객 이탈 예측 모형에 의한 구독 중지 정확율은 평균 51.64%가 된다. 이는 기존에는 평균 6%의 중지 고객을 예측했던 것에 비하여, 이탈 예측 모형을 이용할 경우 평균 50% 이상의 중지 고객을 예측할 수 있다는 것이다.

또한 중지율에 의한 예측은 평균 79.38%의 정확율을 갖는다.

위의 결과로 고객 이탈 예측 모형에 의해 신문 구독 중지 가능 고객을 샘플링 한 뒤 중지 고객을 찾는 방법이 훨씬 생산적이며, 안전적이라는 것을 알 수가 있다.

4. 결 론

본 연구에서는 신문 산업에 알맞은 CRM을 실행하기 위하여, 국내 3대 신문사 중 하나인 A사 고객의 데이터를 이용하여 고객 이탈 예측 모형을 구축하였다.

고객 데이터의 결측 정도를 살펴보기 위해 선택된 데이터 항목에 대해 테이터 기록율을 살펴보았으며, 기록율이 10% 미만으로 낮은 데이터는 탈락시켰다.

또한 고객의 신문 구독 중지 현황에 대해 분석하기 위해 초기에 선택된 102개의 변수와 구독 중지현황에 대한 상관분석을 실시하였다.

이 결과 전체 변수 중 구독 중지에 가장 영향이 있고, 기록율이 정확한 19개 변수를 추출하여 모형을 설계하였다. 그리고 모형의 검증을 위하여 정오분류표, 이득도표, 예측에 의한 정확율 등을 살펴보았다.

신문 산업이나 통신 산업처럼 일정 기간, 일정한 사용료를 지불하고 컨텐츠를 사용하는 산업의 경우, 반복 구매나 고객의 충성도를 향상시키기 위한 고객관계관리 전략을 구축하는데 많은 어려움이 있으나, 본 연구에서는 실 데이터를 이용하여 효과적인 고객 이탈 예측 모형 구축 사례를 제안하였다.

추후 연구 과제로 좀 더 의미 있는 고객 이탈 예측 모형이 되기 위해서는 기본적으로 데이터의 기록율이나 정확도가 뒷받침되어야 한다.

또한 신문 구독 고객의 이탈을 막기 위해서 관련 캠페인, 이벤트와 연계하여 적절한 캠페인 전략을 기획해야 좀 더 큰 효과를 얻을 수가 있을 것이다.

5. 참 고 문 헌

[1] 강현석, 서영호, "고객구매활동 기반의 e-CRM 전략", 품질경영학회지 제28권 제3호 (2000) :133-144

[2] 김충영, 장남식, 김준우, "이동통신서비스 해지 고객 예측 모형의 비교 분석에 관한 연구", 경영정보학연구 제12권 제1호 (2002) :139-158

[3] 송관배, 양광모, 강경식, "CRM 분석을 위한 고객 세분화에 관한 연구", 안전경영과학회지 제5권 제3호 (2003) :133-143

[4] 윤여중, 이상곤, "CRM서 고객정보 관리활동의 매개적 역할에 관한 연구", 한국정보기술응용학회 춘계학술대회 (2006) :345-354

[5] 윤종욱, "CRM에서 고객세분화를 위한 균형적 관점", 한국정보기술응용학회 춘계학술대회 (2004) :300-309

[6] 이현수, 채영일, "고객지식의 획득/활용과 고객관계관리에 영향을 미치는 요인", 경영과학 제22권 제1호 (2005) :127-148

[7] 장일동, "통신시장에서 신경망을 통한 고객관리 분석", 한국OA학회 논문집 제6권 제3호 (2001) :29-34

[8] 정현주, 고준, 김영걸, "고객관계관리에서 고객정보/고객지식 품질에 영향을 미치는 요인 : 서비스 산업을 중심으로", 경영과학 제19권 제2호 (2002) :1-23

[9] 최대우, "전략적 CRM, 현재와 미래", 정보과학회지 제21권 제10호 (2003) :74-78

[10] Apte,C., Weiss,S, "Data mining with decision trees and decision rules", Future Generation Computer Systems 13 (1997) :197-210

[11] HC Kang, ST Han, JH Choi, "Interpretation of Data Mining Prediction Model Using Decision Tree", The Korean Communications in Statistics, Vol.7, No.3 (2000) : 937-943

저 자 소 개

양 승 정

서울산업대학교에서 학사와 석사를 취득하고, 동국대학교 산업공학과 박사졸업예정이며, 현재 동국대학교 산업기술연구원에서 전임연구원으로 재직 중에 있다.

주소: 서울시 중구 필동 26번지 동국대학교 본관 C106호 산업기술연구원

이 종 태

서울대학교에서 학사를, KAIST에서 석사를 마치고 미국 캘리포니아(버클리)대학교 산업공학과에서 공학박사학위를 취득하였으며, 퍼듀대학에서 박사후과정 연구원을 거쳐 현재는 동국대학교 산업공학과 교수로 재직 중에 있다.

주소: 서울시 중구 필동 26번지 동국대학교 원흥관 E416호