Journal of the Korean Data & Information Science Society
2009, **20**(4), 615–627

한국데이터정보과학회지

화장품구매 자료를 통한 고객 구매행태 분석[†]

조대현[1] · 김병수[2] · 석경하[3] · 이종언[4] · 김종성[5] · 김선화[5]

[1235]인제대학교 데이터정보학과 · [4]미켄즈(주)

접수 2009년 1월 26일, 수정 2009년 5월 28일, 게재확정 2009년 6월 10일

요 약

본 연구의 목적은 효과적인 마케팅전략 수립에 도움이 되는 정보를 제공하는 데 있다. 이를 위하여 화장품구매 자료로부터 고객 구매형태와 재구매 간의 관계를 분석하여 고객충성도 예측모형을 개발하였다. 고객충성도는 재구매 가능성으로 측정하였다.

본 연구에서 사용된 자료는 국내의 한 화장품회사 고객들의 2000년부터 2008년까지 9년간의 구매자료 (432,528명, 2,440,107건)이다. 예측모형의 목표변수는 재구매 유무이고, 설명변수는 구매수량, 구매액, 휴면기간 등의 기본변수와 구매횟수와 거래 일자를 이용한 가공변수들이다. 충성도 예측모형은 데이터마이닝 기법인 로지스틱회귀, 의사결정나무 및 신경망모형을 사용하였다. 예측모형평가의 측도로는 하이드게 점수를 사용하였으며, 최대의 하이드게 점수를 가지는 분계점을 선택하였다. 각 예측모형에서 선택된 변수는 유사하며, 모형비교 결과 세 모형의 효율과 평가측도의 차이는 크지 않았다. 정분류율이 다소 높고 해석과 활용이 쉬운 의사결정나무모형을 최종모형으로 선택했다.

주요용어: 고객 충성도, 데이터마이닝, 예측모형, 재구매율, 하이드게 점수.

1. 서론

경제 성장과정에서 제품이 부족했던 과거에는 값싸고 좋은 품질의 공산품을 대량생산하여 공급하는 것이 주관심사였다. 이 시대의 마케팅은 고객의 개별적인 특성을 중요하게 여기지 않고 시장전체를 대상으로 하는 매스마케팅이 중심이었다.

그러나 오늘날 시장은 네트워크의 발달과 기업간의 기술 평준화로 인해 경쟁사보다 기술의 우위로 시장을 석권하기가 점점 어려워지고 있다. 또한 경쟁사의 난립으로 인해 제품과 서비스가 각 고객의 개성을 만족시켜주지 못한다면 고객들의 이탈을 막기 어렵게 되었다.

정보기술의 발전으로 고객관련 정보들과 거래실적 등을 데이터베이스로 구축하게 됨으로써 고객성향, 구매실적, 기업에 대한 기여도 등을 분석할 수 있는 기반이 형성되었고, 고객의 구매행태 분석을 통하여 고객에 대한 차별적인 마케팅을 수행하는 것이 가능하게 되었다. 이러한 기업 환경과 사회의 변화에 따라 고객관계관리 (CRM: Customer Relationship Management) 시스템이 중요한 마케팅 기법으로 자리 잡게 되었다 (백신정, 2004; Judy와 Raymond, 2001; Cho와 Park, 2008). Mercer 컨설팅사의 보고

[†] 본 연구는 산학협동재단 2008년 학술연구비 지원에 의해 이루어졌음.
[1] (621-749) 경남 김해시 어방동 607, 인제대학교 데이터정보학과/통계정보연구소, 교수.
[2] 교신저자: (621-749) 경남 김해시 어방동 607, 인제대학교 데이터정보학과/통계정보연구소, 조교수.
E-mail: kbs@stat.inje.ac.kr
[3] (621-749) 경남 김해시 어방동 607, 인제대학교 데이터정보학과/통계정보연구소, 교수.
[4] (135-010) 서울시 강남구 논현동 206, 일양빌딩 7층 미켄즈(주), 연구소장.
[5] (621-749) 경남 김해시 어방동 607, 인제대학교 데이터정보학과, 석사과정.

서에 따르면 최고경영자의 66%가 고객과의 관계구축 및 유지가 21세기 기업의 최고 경쟁력이 될 것으로 예상하고 있다. 이는 CRM의 중요성을 설명해주는 부분이기도 하다.

CRM 활동의 구체적 목표는 고객정보의 체계적 분석과 이에 근거한 영업 및 마케팅 활용시스템의 구축을 통해 기존고객의 유지, 신규고객 확보, 고객의 평생가치 극대화로 나눌 수 있다 (당현준, 2003; Ko와 Lee, 2006). 기존고객과의 좋은 관계유지를 통해 얻을 수 있는 이익이 신규고객으로부터 얻는 이익에 비해 매우 크므로 CRM의 주된 대상은 기존고객이라고 할 수 있다. 고객유지를 위해서는 고객정보 혹은 고객의 구매행태로부터 고객들을 잘 파악하여 그에 맞는 마케팅 적용이 필요하다. 현재 고객이 미래에 다시 제품을 구매할 가능성을 고객의 충성도라고 정의할 때 고객의 충성도는 고객의 구매행태나 인구통계학적 특성에 따라 결정될 것이다. 기업이 각 고객의 충성도인 재구매 가능성 여부에 따른 차별화된 마케팅 전략을 사용함으로써 비용을 줄임은 물론 보다 좋은 판매효과를 통한 이익의 극대화를 기대해 볼 수 있을 것이다.

고객 데이터 분석을 위한 중요 기술 중 하나가 데이터마이닝 기법이다. 이 기법을 통하여 기업이 보유하고 있는 일일 거래자료, 고객자료, 상품자료, 마케팅활동의 피드백 자료와 기타 외부자료를 포함하여 사용 가능한 데이터를 기반으로 숨겨진 지식, 패턴, 법칙과 관계를 발견하고 이를 실제 경영에서 의사결정을 위한 정보로 활용하고자 하는 것이다 (백신정, 2004). 또 다른 데이터마이닝과 CRM에 관한 연구로는 전성해 등 (2008)과 이도현 (2000) 등이 있다.

본 연구에서는 화장품회사의 고객 구매자료를 이용하여 제품을 한번 혹은 그 이상 구매를 일으킨 기존고객을 대상으로 고객 구매행태와 재구매 간의 관계를 분석하여 얻어진 결과, 재구매 가능성 즉, 충성도를 예측하는 충성도 예측모형을 개발하여 효과적인 마케팅 수립에 도움이 되는 정보를 제공하고자 한다.

2절에서는 자료 탐색을 통한 자료에 대한 이해와 재구매에 영향을 주는 새로운 변수들을 생성하고, 기존 변수와 생성변수들에 대한 탐색을 다룬다. 3절에서는 종속변수가 이진인 경우 전형적인 분석모형인 의사결정나무 분석기법 (최종후 등, 1998)과 로지스틱회귀모형 (김순귀 등, 2003) 및 신경망모형을 적합하여 분석한 결과 및 이들 모형들에 대한 비교를 다루며, 4절은 결론으로 이루어진다.

2. 자료탐색

2.1. 원 자료

원 자료는 2000년부터 2008년까지 9년간의 구매정보로 고객 구매자료 (432,528명, 2,440,107건)와 제품에 대한 정보를 나타내는 제품자료로 구성되어있다. 분석에 사용된 자료는 고객 구매자료를 기준으로 하되 제품자료로부터 고객이 구매한 제품품목에 대한 정보를 추가하였다. 분석용 자료의 변수는 크게 고객 신상정보와 구매정보로 나누어지며 표 2.1과 같다.

2.2. 분석자료와 변수생성

2.2.1. 분석자료

원 자료로부터 분석에 사용할 자료와 변수를 생성하였다. 원 자료에서 구매액이 음의 값을 갖는 경우는 환불 혹은 오기로 판단하여 제거하였다. 또한 고객을 기준으로 평균 구매수량, 총 구매수량, 평균 구매금액, 총 구매금액이 각 변수의 상위 0.5%인 8.80, 63개, 980,000원, 6,963,000원을 초과하는 조건 중 하나라도 만족하게 되면 비정상적인 구매로 판단하여 제거하였다. 제거된 결과 제거된 고객 수는 2,431명이고 남은 고객 수는 430,076로 전체고객 수의 약 0.56%가 제거되었다. 원 자료로부터 분석에 사용할 자료와 변수를 생성하였다. 분석에 사용할 자료는 기준시점을 정하고, 기준시점 이전 3년간의 고객들을 대상으로 구매행태 및 나이와 기준시점 이후 1년간의 재구매 여부를 나타내는 변수로 구

표 2.1 원 자료의 변수

구분	변수명	설명
신상정보	cust_id	고객id
	job_code	직업코드 (10, 50, 99 등)
	birth	주민번호 앞4자리 (출처: 고객구매자료)
	sex	성별 (1=남, 2=여, 5=외국인 남, 6=외국인 여)
	marry	결혼여부 (1:기타, 2:기혼, 3:미혼)
	region	거주지역 (1=결측, 2=부산, 3=경기, 4=광주, 5=인천, 6=울산, 7=전북, 8=충북, 9=경북, 10=강원, 11=대구, 12=대전, 13=제주, 14=서울, 15=충남, 16=경남, 17=전남)
구매정보	trade_id	구매id
	date	구매일 (2000.02.22 ∼ 2008.05.20)
	store_code	판매점 코드
	prod_id	제품id
	category	제품 category (Basic, Foundation, Anti-Aging, Whitening)
	trade_nm	구매수량
	amount	구매액 (원)

성하였다. 기준시점은 화장품의 특성상 계절효과를 고려한 2004년 6월 30일부터 2007년 3월 30일까지 3개월 단위로 정하였으며, 전부 12개의 시점, 각 분기당 3개의 시점이다. 12개의 기준시점 각각에 따라 만들어진 자료를 합하여 구성한 분석자료의 자료 수는 2,526,550건으로 총 구매횟수가 1인 고객에 대한 자료의 수가 1,451,264명이고, 총 구매횟수가 2회 이상인 고객에 대한 자료의 수는 1,075,286건이다. 기준시점 전 3년의 구매행태를 본 이유는, 3년간의 구매행태를 기반으로 하여 고객을 충성, 휴면 등으로 분류하고 재구매를 예측하기 위해서이다.

2.2.2. 변수생성

하나의 기준시점의 자료에서 한 고객의 자료는 하나의 관측값을 나타내도록 요약하였으며, 요약 변수를 이용하여 변수를 생성하였고 결과는 표 2.2와 같다.

표 2.2 분석자료에서 고객들의 구매정보와 재구매에 대한 변수

변수명	설명
re	기준시점 이후 1년 동안의 재구매 여부(0:비구매 1:재구매)
total_count	총 구매횟수
min_date	최초 구매일
max_date	마지막 구매일
sum_trade_nm	총 구매수량
mean_trade_nm	평균구매수량 (= sum_trade_nm / total_count)
sum_aount	총 구매액 (단위:원)
mean_amount	평균구매액 (= sum_amount / total_count (단위:원))
cy	구매주기 = (max_date - min_date) / (total_count - 1)
distance	휴면기간 = (기준시점 - max_date + 1)
age	구매당시 나이
re_mean	1년 전 기준시점에서의 고객들의 재구매율
sum_basic	총 Basic 구매수량
sum_whitening	총 Whitening 구매수량
sum_anti_aging	총 Anti_aging 구매수량
sum_foundation	총 Foundation 구매수량
f_basic	마지막 구매 시 Basic 구매수량
f_whitening	마지막 구매 시 Whitening 구매수량
f_anti_agingf_foundation	마지막 구매 시 Anit_aging 구매수량
f_foundation	마지막 구매 시 Foundation 구매수량

목표변수는 기준시점 이후 1년 동안의 구매여부를 나타내는 재구매이며, 재구매한 경우에 1, 재구매하지 않은 경우에 0의 값을 부여하였다. 입력 변수는 총 구매수량, 평균 구매수량, 총구매액, 평균구매액, 기준시점과 마지막 거래일자와의 거리 (휴면기간, distance), 구매횟수 (total_count), 현재 기준시점에서의 1년 전 기준시점의 재구매율 (re_mean), 제품의 범주 (기초, 미백, 노화방지, 파운데이션)별로 총 구매수량 (sum_basic, sum_whitening, sum_anti_aging, sum_foundaion)과 마지막 구매 시 구매수량 (f_basic, f_whitening, f_anti_aging, f_foundation)등이다.

또한 군집분석을 통하여 지역변수를 생성하였다. 지역변수 (region)는 17개의 지역으로 구분되어 있는 범주형 변수이다. 지역의 경제적 특성과 지리적 특성 등으로 구매형태가 다를 것으로 예상하여 비슷한 특성을 지닌 지역끼리 묶어 분석을 용이하게 하고자 군집분석 (허명회 등, 2007)을 이용해 항목을 줄였으며 구매형태에 관련된 평균 구매수량, 평균구매액, 총구매액, 나이 및 ratio가 사용되었다. ratio는 지역별 총 인구 중 제품을 구매한 인구가 얼마나 되는지에 대한 비를 나타내며 지역이 결측값을 갖는 경우는 "*ratio=결측을 제외한 구매고객 수/전체 인구 수* "로 계산하였으며, 지역별 총 인구 수는 2008년 통계청 홈페이지에서 제공한 자료를 이용하였다. 표 2.3은 Ward 방법을 사용한 계층적 군집분석을 이용한 결과를 보여주고 있다. 17개 지역이 4개의 군집으로 형성되었고, 1군집과 2군집의 재구매율은 각각 0.241, 0.243으로 3군집과 4군집보다 높음을 알 수 있다.

표 2.3 군집분석 결과

G_region	구성	재구매율
1	결측(1)	0.241
2	강원(10), 제주(13), 충북(8), 충남(15), 전남(17)	0.243
3	광주(4), 인천(5), 경남(16), 전북(7), 경기(3), 경북(9)	0.209
4	부산(2), 대전(12), 서울(14), 울산(6), 대구(11)	0.228

2.3. 재구매율의 특징

3년간의 구매형태를 기반으로 고객을 분류하고 재구매를 예측하기 위하여 기준시점으로부터 지난 3년간의 구매자료를 이용하여 다음 1년 동안 재구매가 일어난 비율인 재구매율의 월별변화를 살펴본 결과 표 2.4와 같다. 기준시점에 따라 재구매율의 변화를 보면 시간이 지남에 따라 고객의 수는 증가하지만 재구매율은 줄어들고 있음을 알 수 있다.

표 2.4 기준시점에 따른 3개월간 재구매율의 변화 (전체)

	기준시점												
	2004 06-30	2004 09-30	2004 12-31	2005 03-31	2005 06-30	2005 09-30	2005 12-31	2006 03-31	2006 06-30	2006 09-30	2006 12-31	2007 03-31	총합
재구매안함 (0)	69009 (69.77)	81261 (70.10)	97890 (72.95)	117438 (73.08)	132787 (72.77)	151920 (73.21)	174432 (75.92)	195896 (77.31)	218238 (79.69)	237152 (82.05)	237372 (82.14)	238262 (81.65)	1951657 (77.25)
재구매함 (1)	29898 (30.23)	34661 (29.90)	36292 (27.05)	43262 (26.92)	49691 (27.23)	55597 (26.79)	55321 (24.08)	57502 (22.69)	55616 (20.31)	51895 (17.95)	51608 (17.86)	53550 (18.35)	574893 (22.75)
총합	98907 (3.91)	115922 (4.59)	134182 (5.31)	160700 (6.36)	182478 (7.22)	207517 (8.21)	229753 (9.09)	253398 (10.03)	273854 (10.84)	289047 (11.44)	288980 (11.44)	291812 (11.55)	2526550 (100.00)

총 구매횟수가 1인 고객과 총 구매횟수가 2이상인 고객으로 분류하여 재구매율을 살펴보면 표 2.5, 표 2.6과 같다. 구매횟수가 1인 고객은 평균 재구매율이 13.45%인 반면 2 이상 구매고객의 평균 재구매율은 35.31%로 구매횟수가 1인 고객의 2배 이상으로 나타나고 있다. 기준시점별로 구매횟수를 살펴보면

두 경우 모두 재구매율이 낮아지고 있지만 구매횟수가 2회 이상인 고객의 재구매율 감소율이 상대적으로 낮은 것을 알 수 있다.

표 2.5 기준시점에 따른 3개월간 재구매율의 변화 (구매횟수가 1회인 고객)

	기준시점												
	2004 06-30	2004 09-30	2004 12-31	2005 03-31	2005 06-30	2005 09-30	2005 12-31	2006 03-31	2006 06-30	2006 09-30	2006 12-31	2007 03-31	총합
재구매안함 (0)	47132 (79.42)	55366 (80.19)	64572 (83.21)	78231 (82.35)	88778 (82.30)	101338 (82.79)	113908 (85.83)	126567 (86.69)	138431 (88.90)	148037 (90.88)	146717 (91.28)	146983 (90.78)	1256060 (86.55)
재구매함 (1)	12214 (20.58)	13681 (19.81)	13031 (16.79)	16769 (17.65)	19099 (17.70)	21061 (17.21)	18799 (14.17)	19435 (13.31)	17293 (11.10)	14864 (9.12)	14023 (8.72)	14935 (9.22)	195204 (13.45)
총합	59346 (4.09)	69047 (4.76)	77603 (5.35)	95000 (6.55)	107877 (7.43)	122399 (8.43)	132707 (9.14)	146002 (10.06)	155724 (10.73)	162901 (11.22)	160740 (11.08)	161918 (11.16)	1451264 (100.00)

표 2.6 기준시점에 따른 3개월간 재구매율의 변화 (구매횟수가 2회 이상인 고객)

	기준시점												
	2004 06-30	2004 09-30	2004 12-31	2005 03-31	2005 06-30	2005 09-30	2005 12-31	2006 03-31	2006 06-30	2006 09-30	2006 12-31	2007 03-31	총합
재구매안함 (0)	21877 (55.30)	25895 (55.24)	33318 (58.89)	39207 (59.68)	44009 (58.99)	50582 (59.43)	60524 (62.37)	69329 (64.55)	79807 (67.56)	89115 (70.64)	90655 (70.69)	91279 (70.27)	695597 (64.69)
재구매함 (1)	17684 (44.70)	20980 (44.76)	23261 (41.11)	26493 (40.32)	30592 (41.01)	34536 (40.57)	36522 (37.63)	38067 (35.45)	38323 (32.44)	37031 (29.36)	37585 (29.31)	38615 (29.73)	379689 (35.31)
총합	39561 (3.68)	46875 (4.36)	56579 (5.26)	65700 (6.11)	74601 (6.94)	85118 (7.92)	97046 (9.03)	107396 (9.99)	118130 (10.99)	126146 (11.73)	128240 (11.93)	129894 (12.08)	1075286 (100.00)

2.4. 분석변수들의 기술통계량

표 2.7은 전체 자료에서 분석변수들의 기술통계량이다. 평균 구매액의 평균은 약 225,686.56원으로 나타났으며 총 구매액의 평균은 약 507,714.59원으로 나타났다. 제품별 총 구매수량의 평균은 Basic이 2.78로 가장 높고 제품별 마지막 구매 시 구매수량 또한 Basic이 1.03으로 가장 높게 나타났다. 평균 휴면기간은 약 355.51일이며 평균 구매기간은 약 123.57일이다.

2.5. 재구매 유무에 따른 변수들의 평균비교

표 2.8은 구매회수가 1인 경우와 2 이상인 경우로 나누어 재구매 유무에 따른 평균비교를 한 결과이다. 모든 변수들의 p-value가 0.001 미만이므로 고려된 변수들이 재구매와 관계가 깊은 것으로 판단되며, distance와 cy를 제외한 모든 변수들은 재구매와 양의 관계를 보이고 있다.

2.6. 상관분석

표 2.9는 전체 고객들에 대한 변수들의 상관분석결과이다. 상관관계가 높은 변수들로는 mean_amount와 mean_trade_nm, sum_amount와 sum_trade_nm, total_count와 sum_tradenm, 그리고 total_count와 sum_amount 등이며 재구매여부와 상관관계가 가장 높은 변수는 distance이다.

표 2.7 분석변수들의 기술통계량

변수	N	평균값	표준편차	최소값	최대값
mean_trade_nm	1075286	2.10	1.07	1	29.5
mean_amount	1075286	225686.56	122981.76	50000	2683500
sum_trade_nm	2526550	4.72	5.75	1	63
sum_amount	2526550	507714.59	627190.91	50000	6962000
age	1869635	33.13	8.17	9	91
total_count	2526550	2.27	2.42	1	42
sum_basic	2526204	2.78	3.87	0	63
sum_whitening	2526204	0.91	1.69	0	42
sum_anti_aging	2526204	0.48	0.94	0	27
sum_foundation	2526204	0.54	1.18	0	41
f_basic	1074419	1.03	1.16	0	24
f_whitening	1074419	0.37	0.70	0	30
f_anti_aging	1074419	0.21	0.47	0	20
f_foundation	1074419	0.26	0.53	0	33
distance	2526550	355.51	288.65	0	1095
cy	1075286	123.57	122.25	1.	1092
re_mean	2526550	0.27	0.12	0.13	0.45

표 2.8 재구매 유무에 따른 평균비교

			평균		Satterthwaite's	
총 구매횟수	N	변수	0	1	t Value	$Pr > \|t\|$
1회구매고객	1451264	sum_trade_nm	1.99	2.26	-63.31	< .0001
	(57.44%)	sum_amount	215586.45	245084.49	-59.68	< .0001
		age	32.85	33.64	-35.49	< .0001
		sum_basic	1.16	1.28	-35.30	< .0001
		sum_whitening	0.39	0.43	-22.41	< .0001
		sum_anti_aging	0.21	0.25	-32.29	< .0001
		sum_foundation	0.24	0.29	-43.34	< .0001
		distance	442.81	219.77	378.69	< .0001
		re_mean	0.17	0.18	-113.78	< .0001
2회이상구매고객	1075286	mean_trade_nm	2.06	2.16	-46.01	< .0001
	(42.56%)	mean_amount	222225.66	232026.97	-40.20	< .0001
		sum_trade_nm	7.18	10.48	-214.94	< .0001
		sum_amount	773070.45	1122997.52	-208.20	< .0001
		age	32.92	33.88	-53.71	< .0001
		total_count	3.46	4.94	-229.03	< .0001
		sum_basic	4.31	6.13	-172.66	< .0001
		sum_whitening	1.38	1.98	-121.98	< .0001
		sum_anti_aging	0.70	1.08	-139.90	< .0001
		sum_foundation	0.78	1.25	-132.88	< .0001
		f_basic	1.01	1.07	-26.64	< .0001
		f_whitening	0.36	0.39	-19.74	< .0001
		f_anti_aging	0.20	0.23	-37.05	< .0001
		f_foundation	0.25	0.29	-35.70	< .0001
		distance	356.53	134.62	541.87	< .0001
		cy	130.09	111.63	79.38	< .0001
		re_mean	0.40	0.41	-92.21	< .0001

3. 재구매 예측모형

분석자료에 포함된 변수들을 이용하여 재구매 유무를 예측할 수 있는 모형을 개발하였다. 구매횟수가 1회인 경우 변수 생성과정에서 구매행태에 대한 변수가 결측이 많이 나타나고 있어 좋은 예측모형을 기

표 2.9 전체 고객들에 대한 변수들의 상관분석

	mean_ trade_nm	mean_ amount	sum_ trade_nm	sum_ amount	age	max_ count	distance	cy	re_mean
mean_amount	**0.95**	.	.	.	.	.	.	.	.
sum_trade_nm	**0.46**	**0.44**	.	.	.	.	.	.	.
sum_amount	**0.46**	**0.49**	**0.99**	.	.	.	.	.	
age	-0.01	0.01	0.02	0.02	.	.	.	.	.
total_count	0.00	0.00	**0.87**	**0.85**	0.02	.	.	.	.
distance	-0.02	-0.02	-0.23	-0.22	-0.07	-0.25	.	.	.
cy	-0.06	-0.05	-0.21	-0.21	-0.04	-0.22	-0.10	.	.
re_mean	0.02	0.02	**0.53**	**0.52**	0.01	**0.59**	-0.26	-0.07	.
re	0.04	0.04	0.28	0.27	0.05	**0.30**	**-0.36**	-0.07	0.27

대할 수 없으므로 총 구매횟수가 2회 이상인 고객들을 대상으로 예측모형을 개발하기로 한다. 예측모형으로는 의사결정나무모형과 로지스틱회귀모형 및 신경망모형을 사용하였다. 전체 자료를 훈련용 40%, 검증용 30%, 평가용 30%로 분할하여 모형의 훈련과 검증 및 모형비교를 위한 평가용으로 활용하였다.

각 모형에서 재구매 확률 또는 점수를 계산하고 주어진분계점에 따라 표 3.1과 같이 정오분류표를 작성하여 예측성평가를 하였다. 본 연구에서는 Sohn과 Lee (2006)에서 언급한 하이드게 점수인 HSS (Heidke, 1926)가 최대가 되도록 분계점을 정하였다. HSS는 다음과 같은 식으로 정의된다.

$$HSS = \frac{PCM - PCR}{1 - PCR},$$

여기서 $PCM = (A + D)/N$ 은 예측모형에 의해 정확히 예측되는 확률(정분류율)에 해당되고, $PCR = (O_1F_1 + O_2F_2)/N^2$ 은 임의예측에 의해 정확히 예측되는 확률에 해당되는 값이다. HSS는 모형에 의한 예측이 임의예측보다 나은 정도를 나타내는 값이다. 본 연구에서는 검증용 자료를 사용하여 HSS를 최대로 하는 분계점을 선택하였다.

표 3.1 총 범주가 2인 예측모형에서의 정오분류표

		예측		
		0	1	총계
관측값	0	A	B	$O_1 = A + B$
	1	C	D	$O_2 = C + D$
총계		$F_1 = A + C$	$F_2 = B + D$	$N = A + B + C + D$

3.1. 의사결정나무모형

의사결정나무모형 구축을 위해서 분리기준으로 카이제곱 통계량과 엔트로피 지수 및 지니지수를 사용해 본 결과 정분류율이 각각 74.43%, 74.02%, 74.43%로서 카이제곱 통계량과 지니지수인 경우 같게 나왔으며 엔트로피지수를 사용할 경우 다소 나쁘게 나타났다. 이를 토대로 보편적으로 사용되고 있는 카이제곱 통계량을 분리기준으로 사용하였으며 분리를 위한 유의수준은 0.2로 하였다.

그림 3.1은 분석결과의 나무구조를 나타내며, distance, total_acount, cy가 재구매를 예측하는데 중요한 변수임을 알 수 있다. 훈련용 자료, 검증용 자료에서 HSS를 최대로 하는 분계점은 모두 0.39로 동일했으며, 정분류율은 훈련용 자료에서 약 74.42%, 검증용 자료의 분계점을 토대로 한 평가용 자료에서 약 74.43%였다. 평가용 자료에 대한 재구매 유무 예측결과는 표 3.2와 같다.

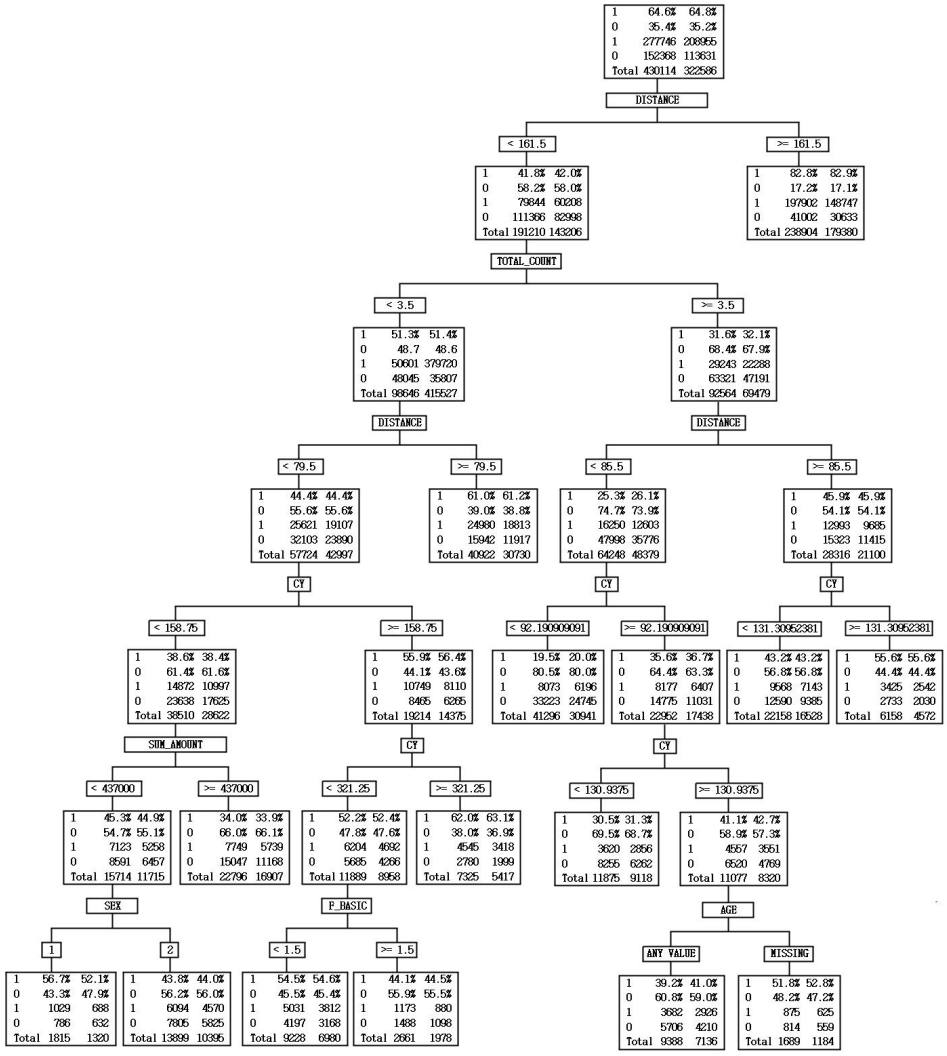

그림 3.1 의사결정나무모형

표 3.2 의사결정나무모형의 정오분류표

		예측		
		0	1	총계
관측값	0	170788	38108	208896
		(52.94)	(11.81)	(64.76)
	1	44386	69304	113690
		(13.76)	(21.48)	(35.24)
총계		215174	107412	322586
		(66.70)	(33.30)	(100.00)

3.2. 로지스틱회귀모형

각 변수와 재구매와의 관계를 선형화를 위해 표 3.3과 같이 변수를 생성하였다. 휴면기간은 지수함수로 적합하였으며 총 구매수량은 2차함수로, 구매횟수와 나이는 각각 지수함수와 4차함수로 적합하였다. 구매주기는 다른 변수와 달리 두 부분으로 나눠지는 것을 그림을 통해 확인하였고, 1차와 3차로 기준을 달리하여 적합 한 뒤 SSE의 합이 가장 적은 40.5세를 기준으로 40세보다 작거나 같은 쪽은 1차로 적합하고 40세보다 큰 쪽은 3차로 적합하였다.

표 3.3 선형화를 위한 새로운 변수

변수명	설명	x
p_distance	$-2.9765 + 3.9926e^{-0.00298\times x}$	distance
p_sum_trade_nm	$-1.4739 + 0.1295x - 0.00201x^2$	sum_trade_nm
p_total_count	$0.9535 - 2.9648e^{-0.1811x}$	total_count
p_age	$3.1079 + 0.000001304x^4 - 0.00024x^3 + 0.0155x^2 - 0.4053x$	age
p_cy	$40 \le cy : -1.4946 + 0.0313x$ $40 > cy : -0.0114 - 0.00546x + 0.000007517x^2 - 0.00000000288x^3$	cy

변수선택을 통하여 로지스틱회귀모형에서 최종적으로 사용된 설명변수는 표 3.3에서 설명한 새로운 변수를 포함한 18 개 변수이다. 로지스틱회귀모형분석 결과 표 3.4와 같은 회귀계수를 얻었으며, 다음과 같은 로지스틱회귀모형으로 표현된다.

$$E(Y) = P(Y = 1) = p$$
$$Logit(p) = \log(p/(1-p)) = -0.9598 + 0.0432\text{f\_anti\_aging} + \cdots + 0.4473\text{p\_total\_acount}$$

여기서 p 는 각각의 고객에 대한 재구매 확률을 의미한다. Wald- χ^2 값을 보면 p_distance와 re_mean이 각각 44190.71, 1278.69로 큰 값을 가지므로 재구매 예측에 영향을 많이 주는 변수라 할 수 있다.

표 3.4 로지스틱회귀모형에서 회귀계수 추정치

Parameter	DF	Estimate	s.e	Wald- χ^2	p-값
Intercept	1	-0.9598	0.0735	170.59	$< .0001$
f_anti_aging	1	0.0432	0.0087	24.57	$< .0001$
f_basic	1	0.0845	0.0043	378.37	$< .0001$
f_foudnation	1	0.0646	0.0098	43.22	$< .0001$
f_whitening	1	0.0492	0.0063	60.36	$< .0001$
g_region 1	1	0.0241	0.0211	1.31	$< .0001$
g_region 2	1	0.0106	0.0089	1.41	0.2525
g_region 3	1	-0.0684	0.0093	54.54	0.2352
mean_amount	1	7.351×10^{-7}	1.329×10^{-7}	30.59	$< .0001$
mean_trade_nm	1	-0.1090	0.0174	39.14	$< .0001$
p_age	1	0.7547	0.0380	394.21	$< .0001$
p_cy	1	0.4374	0.0150	850.57	$< .0001$
p_distance	1	0.8741	0.0042	44190.71	$< .0001$
p_sum_trade_nm	1	0.3973	0.0369	115.72	$< .0001$
re_mean	1	5.2254	0.1461	1278.69	$< .0001$
sex	1	-0.0671	0.0229	8.61	0.0033
sum_amount	1	-1.8×10^{-7}	1.266×10^{-8}	201.08	$< .0001$
sum_foundation	1	0.0323	0.0035	86.74	$< .0001$
p_total_acount	1	0.4473	0.0294	231.95	$< .0001$

훈련용 자료와 검증용 자료에서의 HSS를 최대로 하는 분계점은 각각 0.44와 0.43이었으며, 훈련용

자료에서 정분류율은 73.76%였으며, 검증용 자료의 분계점을 토대로 한 평가용 자료에 대한 재구매 유무를 예측한 결과는 표 3.5의 정오분류표와 같고 정분류율은 73.64%이다.

표 3.5 로지스틱회귀모형의 정오분류표

		예측		
		0	1	총계
관측값	0	166225	42671	208896
		(51.53)	(13.23)	(64.76)
	1	42358	71332	113690
		(13.13)	(22.11)	(35.24)
총계		208583	114003	322586
		(64.66)	(35.34)	(100.00)

3.3. 신경망 모형

신경망모형의 구조는 하나의 입력층과 하나의 은닉층 그리고 하나의 출력층으로 구성하였으며 활성함수 (activation function)는 Hyperbolic Tangent함수를 사용하였다. 은닉층의 뉴런수를 결정하기 위해 여러 뉴런 수에 대해 신경망에 적합시킨 결과는 표 3.6과 같으며, SBC (Schwarz Bayesian Criterion)의 값이 가장 작은 4로 하였다.

표 3.6 뉴런 수에 따른 신경망모형의 비교

뉴런수	Root ASE	Vaild Root ASE	Test Root ASE	Schwarz Bayesian Criterion
1	0.43185	0.43143	0.43165	478645.12
2	0.43057	0.43021	0.43039	476283.29
3	0.42996	0.42961	0.42984	475437.35
4	0.42964	0.42926	0.42951	**475044.95**
5	0.43001	0.42968	0.42990	476122.13

최종 신경망모형에 대한 적합결과 훈련용자료, 검증용자료에서 HSS를 최대로 하는 분계점은 각각 0.42, 0.41이며, 훈련용자료에서 정분류율 약 74.11%, 검증용자료의 분계점을 토대로 한 평가용자료에서의 정분류율은 약 73.94%로 나타났다. 평가용자료에 대한 재구매 유무를 예측한 결과 정오분류표는 표 3.7과 같다.

표 3.7 신경망모형의 정오분류표

		예측		
		0	1	총계
관측값	0	168325	40571	208896
		(52.18)	(19.42)	(64.76)
	1	43507	70183	113690
		(13.49)	(21.76)	(35.24)
총계		211832	110754	322586
		(65.67)	(34.33)	(100.00)

3.4. 모형비교 및 최종모형 선택

앞의 3가지 모형 즉, 의사결정나무, 로지스틱회귀모형, 신경망모형의 결과에 대한 비교는 표 3.8과 같다. 각 모형의 HSS를 최대로 하는 검증용자료의 분계점은 각각 0.40, 0.43 그리고 0.40으로 의사결정나

무모형과 신경망 모형이 같으며 로지스틱회귀모형이 약간 높게 나타나고 있지만 크게 차이가 나지 않게 선택되었다. 세 모형의 평가용자료로부터 제곱근평균제곱오차 (Root ASE)를 비교해 보면 의사결정나무모형이 0.4201로 로지스틱회귀모형 (0.4315)와 신경망모형 (0.4303)보다 작으므로 조금 우수한 것으로 나오며, 평가용자료의 정분류율 또한 의사결정나무모형에서 약74.43%로 다른 두 모형 보다 우수한 것으로 나타났다. 신경망모형의 성능은 각 측도에서 중간을 차지하고 있다. 따라서 선택된 최종모형은 제곱근 평균제곱오차가 작고 정분류율이 높으면서 해석과 이해가 용이한 점이 장점인 의사결정나무 모형이다.

표 3.8 의사결정나무, 로지스틱회귀, 신경망모형의 비교

모형	분계점*	HSS*	HSS**	정분류율(%)**	Root ASE**
의사결정나무모형	0.40	0.4319	0.4326	74.43	0.4201
로지스틱회귀모형	0.43	0.4224	0.4229	73.64	0.4315
신경망모형	0.41	0.4253	0.4256	73.94	0.4295

*검증용자료, **평가용자료

4. 결론

본 연구는 화장품 구매자료를 이용하여 과거 3년의 구매행태를 분석하여 이를 바탕으로 향후 1년 이내의 재구매 여부를 예측하는 모형을 개발하기 위해 의사결정나무모형과 로지스틱회귀모형 및 신경망모형을 이용하였다.

모형개발을 위해 사용된 자료는 총 9년간의 고객 구매자료이며, 이용된 변수는 표 2.2와 같다. 각 변수들에 대한 특징을 살펴보고, 재구매 유무에 따른 평균비교와 상관분석을 통하여 재구매에 유의한 영향을 주는 변수를 살펴보았다. 로지스틱회귀모형과 신경망 모형을 위해 그래프를 이용하여 선형화 한 변수를 사용하였으며 세 모형에서 재구매에 유의한 영향을 주는 변수는 비슷한 것으로 나타났다.

각 모형에서 재구매에 대한 점수를 계산하고 최적의 분계점은 찾기 위해 HSS를 이용하였으며, 검증용자료에서 HSS 값을 최대로 하는 분계점을 기준으로 평가용자료에 적용하여 정분류표를 작성하였고, 예측성 평가를 하였다. 세 모형의 분계점을 의사결정나무모형과 신경망 모형이 0.40으로 동일하며 로지스틱 모형이 0.43으로 나타났다.

세 모형을 비교해 보면 HSS는 검증용자료와 평가용자료 모두 의사결정나무모형이 높게 나왔으며, 검증용자료의 분계점을 기준으로 평가용자료의 정분류율은 의사결정나무모형이 약 74.43%로 가장 높고, 다음으로 신경망모형이 약 73.74%로 나왔고 로지스틱회귀모형이 약 73.64%로 가장 낮게 나타났다. 각 모형별로 재구매자를 재구매자로 분류하는 비율은 신경망모형이 우수하게 나타났다. 세 모형비교에서 결과적으로 각 측도간의 차이는 크지 않게 나타나지만 의사결정나무모형이 다소 우수한 것으로 나타났고 해석과 활용이 쉬운 모형이므로 최종모형으로 선택하였다.

참고문헌

김순귀, 정동빈, 박영술 (2003). <SPSS를 활용한 로지스틱 회귀모형의 이해와 응용>, SPSS 아카데미, 서울.
당현준 (2003). <CRM 을 이용한 출판마케팅 전략연구>, 석사학위논문, 이화여자대학교, 서울.
백신정 (2004). <데이터마이닝을 통한 고객관리데이터의 분석>, 석사학위논문, 고려대학교, 서울.
이도현 (2000). 데이터마이닝을 이용한 CRM, <정보과학회지>, **18,** 4-11.
전성해, 김승화, 전홍석 (2008). <데이터마이닝과 CRM>, 자유아카데미, 서울.
최종후, 한상태, 강현철, 김은석 (1998). <데이터마이닝의 의사결정나무분석>, 고려정보산업, 서울.

허명회, 양경숙 (2007). <SPSS 다변량자료분석>, SPSS 아카데미, 서울.

Cho, M. H. and Park, E. S. (2008). Analyzing customer management data by data mining: Case study on churn prediction models for insurance company in Korea. *Journal of the Korean Data & Information Science Society*, **19**, 1007-1018.

Heidke, P. (1926). Berechnung des Erfolges und der Gute der Windstar kevorhersagen im Sturmwarnungs-dienst. *Geografiska Annaler*, **8**, 301-349.

Strauss, J., Ansary, A. E. and Frost, R. (2001). *E-Marketing,* Prentice Hall.

Ko, B. S. and Lee, S. W. (2006). Customer behavior analysis on mobile advertisement. *Journal of the Korean Data & Information Science Society*, **17**, 1251-1259.

Sohn, K. T. and Lee, E. H. (2006). Guidance on choice of skill score for determination of thresholds in ternary forecast. *Journal of the Korean Data Analysis Society*, **8**, 2553-2565.

Journal of the Korean Data & Information Science Society
2009, **20**(4), 615–627

한국데이터정보과학회지

A study on the behavior of cosmetic customers[†]

Daehyeon Cho[1] · Byungsoo Kim[2] · Kyungha Seok[3] ·
Jongun Lee[4] · Jongsung Kim[5] · Sunhwa Kim[5]

[1235]Department of Data Science, Inje University
[4]Head of research center, Micans

Received 26 January 2009, revised 28 May 2009, accepted 10 June 2009

Abstract

In micro marketing promotion, it is important to know the behavior of customers. In this study we are interested in the forecasting of repurchase of customers from customers' behavior. By analyzing the cosmetic transaction data we derive some variables which play an important role in the knowledge of the customers' behavior and in the modeling of repurchase. As modeling tools we use the decision tree, logistic regression and neural network model. Finally we decide to use the decision tree as a final model since it yields the smallest RASE (root average squared error) and the greatest correct classification rate.

Keywords: Decision tree, logistic regression, micro marketing, neural network, promotion, repurchase.

[†] This work was supported by the 2008 research grant from the The Korea Sanhak Foundation.
[1] Professor, Department of Data Science/Institute of statistical Information, Inje University, Kimhae 621-749, Korea.
[2] Corresponding author: Assistant Professor, Department of Data Science/Institute of statistical Information, Inje University, Kimhae 621-749, Korea. E-mail: kbs@stat.inje.ac.kr
[3] Professor, Department of Data Science/Institute of statistical Information, Inje University, Kimhae 621-749, Korea.
[4] Head of research center, Micans, Ilyang Bulding 7, Nonhyun Dong, Kangnam Gu, Seoul 135-010, Korea.
[5] Graduate student, Department of Data Science, Inje University, Kimhae 621-749, Korea.