理學碩士學位論文

재구매 예측 모형 및 고객 분류

仁濟大學校 大學院

데이터정보學科 데이터정보학專攻

金　宗　性

指導敎授　石　璟　夏

재구매 예측모형 및 고객 분류

仁濟大學校 大學院

데이터정보學科 데이터정보學 專攻

金　宗　性

이 論文을 理學碩士論文으로 提出함

指導敎授　石　璟　夏

2010年　6月

金宗性의 理學碩士學位論文을 認定함.

委員長_____印

委　員_____印

委　員_____印

仁濟大學校 大學院

2010年　6月

목 차

초 록

재구매 예측모형 및 고객 분류

김 종 성

(지도교수 : 석 경 하 교수님)

데이터정보학과 대학원

인제대학교

본 연구의 목적은 효과적인 마케팅전략 수립에 도움이 되는 정보를 제공하는 데 있다. 이를 위하여 화장품구매 자료로부터 고객들의 구매에 관한 거래 기록을 분석하여 고객 구매형태와 재구매 간의 관계를 분석하여 재구매 예측모형을 개발하였다. 재구매 예측모형은 데이터마이닝 기법인 로지스틱회귀, 의사결정나무 및 신경망 모형을 사용하였다. 모형비교 결과 세 모형의 효율과 평가측도의 차이는 크지 않았으나, 정분류율이 다소 높은 신경망모형을 최종모형으로 선택했다. 재구매 예측모형을 통하여 재구매 여부를 예측 하였으며, 재구매 확률을 구할 수 있었다. 성향이 비슷한 고객들을 하나의 그룹으로 묶는다면 효율적인 마케팅전략을 수립 할 수 있다. 이를 위하여 제품을 구매할 확률에 따라 고객을 분류하는 경우와 매출기여도가 높은 순으로 고객을 분류하였다. 개별고객에 대한 향후 재구매 여부를 예측하고, 상황에 맞는 고객 분류를 통해 마케팅전략을 수립하는 방법을 제시한다.

키워드 : 고객 분류, 재구매 예측모형, 재구매율

Abstract

Repurchasing Prediction Model and Customer Classification

Jong Sung Kim

(Director : Prof. Kyung Ha Seok, Ph.D.)

Department of Data Science, Graduate School

Inje University

It is very important to classify customers by analyzing their behaviors with their transaction records. To improve the repurchasing rate some marketing strategies are required. The method includes mining information, modeling of customers' behavior and segmentation,

In this study we develop a prediction model using some derived variables. With the developed model we propose an effective marketing strategies. The proposed method could strengthen the competitive power of companies.

Key words : customer classification, repurchasing prediction model

Ⅰ. 서 론

기업이 추구하는 궁극적인 목적은 이윤창출에 있다고 할 수 있다. 제품이 부족했던 과거에는 이윤창출을 위해 값싸고 좋은 품질의 공산품을 대량생산하여 공급하는 것이 주관심사였다. 그러나 오늘날 기술의 발달과 기업들 간의 기술평준화로 인해 경쟁사보다 기술의 우위만 가지고 이윤창출을 내기 어렵게 되었다. 이로 인해 누구에게 어떻게 제품을 알려서 더 많은 이익을 창출할 수 있을지가 주관심사가 되고 있다.

이러한 목적을 달성하기 위해서 무엇보다 고객과의 관계를 효율적으로 관리하는 것이 필요하며 이러한 기법을 지칭하는 용어가 CRM (Customer Relation Management, 고객관계관리)(김이태, 2005)이다. CRM은 고객들의 성향과 욕구를 미리 파악하여 이를 충족시켜줌으로써 목표로 하는 수익이나 광고효과 등 원하는 바를 얻어내는 기법이다. 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 기업환경과 사회의 변화에 따라 고객관계관리 시스템이 중요한 마케팅 기법으로 자리 잡게 되었다 (백신정, 2004; Judy와 Raymond, 2001; Cho와 Park, 2008).

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). 기존고객의 이탈을 막고 신규고객을 계속 확보한다면 이익이 늘어나는 것은 당연한 것이다. CRM의 주된 대상은 기존고객으로 고객들을 잘 파악하여 그에 맞는 마케팅 활동이 필요하다. 기존고객이 미래에 다시 제품을 구매할 가능성을 예측하여 가능성 정도에 따라 차별화된 마케팅 전략을 사용한다면 마케팅 비용의 절감과 높은 판매효과를 통한 이익의 극대화를 기대해 볼 수 있을 것이다. 고객 개인별로 그 고객에 맞는 차별화된 마케

팅 전략을 세운다면 효과는 상당히 좋을 것이다. 하지만 이는 비용적인 측면에서 고려해볼 때 효율적이지 못할 것이다. 고객 정보의 분석을 통해 성향이 비슷한 고객들을 그룹으로 묶어 그룹별 마케팅 전략을 사용한다면 보다 효율적일 것이다.

고객 정보의 분석을 위한 중요 기술 중 하나가 데이터 마이닝 기법이다. 이 기법을 통하여 사용 가능한 데이터를 기반으로 숨겨진 지식, 패턴, 법칙과 관계를 발견하고 이를 실제 경영에서 의사결정을 위한 정보로 활용하고자 하는 것이다 (백신정, 2004).

본 연구에서는 화장품회사의 구매 자료를 이용하여 구매를 한번이상 한 고객을 대상으로 구매행태와 재구매 간의 관계를 분석하여, 재구매 예측모형을 개발하고, 재구매 확률을 이용하여 고객을 분류하여 효과적인 마케팅 수립에 도움이 되는 정보를 제공하고자 한다.

제 2장에서는 자료탐색을 통한 자료에 대한 이해와 재구매에 영향을 주는 변수를 생성하고, 기존변수와 생성변수에 대한 탐색을 다룬다. 제 3장에서는 의사결정나무 모형과 로지스틱회귀모형 및 신경망모형을 적합하여 분석한 결과 및 이들 모형들의 비교를 다룬다. 제 4장에서는 매출액 순서와, 재구매 예측 모형 중 우수한 모형을 가지고 고객을 분류하는 방법을 제시하고 비교 분석한다. 제 5장에서 결론 및 향후 과제에 대하여 기술하였다.

Ⅱ. 자료 탐색

A. 원자료

분석에 사용된 자료는 2005년 1월부터 2008년 12월까지 총 48개월의 화장품 구매 고객들의 구매정보이다. 고객의 수는 362,432명이고 거래수는 2,224,550건으로 고객 1인당 평균 거래수는 6.14건이다. 주어진 원 구매 자료에 나타난 고객들의 구매정보에 대한 변수들은 <표 2.1>과 같으며, 고객코드, 거래코드, 제품코드, 구매개수 구매금액, 거래일자 등의 변수로 구성되어 있다.

<표 2.1> 원자료에서 고객들의 구매정보에 대한 변수

코드	변 수 명
거래코드	trade_code
고객코드	cust_code
제품코드	trade_item_code
구매개수	trade_num
구매금액	trade_payment
거래일자	trade_date

B. 분석자료와 변수생성

원 자료로부터 분석에 사용할 자료와 변수를 생성하였다. 원자료에서 구매횟수와 구매액이 음의 값을 갖는 경우는 환불 혹은 오기로 판단하여 제거하였다. 제거된 결과 제거된 고객 수는 174명이고 남은 고객 수는 362,258명으로 전체고객 수의 약 0.05%가 제거되었다. 분석에 사용할 자료는 2008년 1월 1일을 기준시점으로 하여,

기준시점 이전 3년간의 구매기록이 있는 고객들을 대상으로 구매행태 변수들 및 기준시점 이후 6개월간의 재구매 여부를 나타내는 변수로 구성하였다. 분석자료의 자료수는 300,509건으로 한 고객의 자료는 하나의 관측값을 나타내도록 요약하였으며, 요약 변수를 이용하여 변수를 생성하였고 결과는 <표 2.2>와 같으며, 목표변수는 기준시점 이후 6개월 동안의 구매여부를 나타내는 재구매이며, 재구매한 경우에 1, 재구매 하지 않은 경우에 0의 값을 부여하였다. 입력 변수는 구매횟수, 총 구매개수, 평균 구매개수, 총 구매액, 평균 구매액, 평균 구매주기, 기준시점과 마지막 거래일자와의 거리 (휴면기간, distance) 등이다.

<표 2.2> 분석자료에서 고객들의 구매정보와 재구매에 대한 변수

변수명	설명
re	재구매 여부
total_count	구매횟수
sum_trade_num	총 구매개수
mean_trade_num	평균 구매갯수
sum_trade_payment	총 구매액 (단위:만원)
mean_trade_payment	평균 구매액 (단위:만원)
min_date	최초 구매일
max_date	마지막 구매일
cy	평균 구매주기
distance	휴면기간 (기준시점 - 마지막 구매일)

C. 변수탐색

1. 재구매율

<표 2.3>에서 보는 바와 같이 전체 300,509명 중 40,817명이 재구매를 하였으며, 재구매율은 13.58% 이다.

<표 2.3> 재구매율

재구매 유무	N	%
재구매 안함(0)	259692	86.42
재구매 함(1)	40817	13.58
총합	300509	100.00

2. 분석변수들의 기술통계량

<표 2.4>는 전체 자료에서 분석변수들의 기술통계량이다. 구매횟수는 평균 약 2.5회, 총 구매개수의 평균은 약 6개였으며, 총 구매액의 평균은 580398원, 평균 구매액의 평균은 약 218215원으로 나타났다. 평균구매주기 평균은 약 132일, 휴면기간의 평균은 약 421일이다.

<표 2.4> 분석변수들의 기술통계량

변수	N	평균값	표준편차	최소값	최대값
구매횟수	300509	2.52	3.54	1	635
총 구매개수	300509	5.86	36.39	1	15301
평균 구매갯수	300509	2.24	1.63	1	135
총 구매액 (단위:만원)	300509	58.04	347.30	0	153417.5
평균 구매액 (단위:만원)	300509	21.82	17.25	0	1125
평균 구매주기	132638	132.18	127.72	1	1082
휴면기간	300509	421.01	293.92	1	1094

3. 재구매 유무에 따른 변수들의 평균비교

재구매 유무에 따른 변수들의 평균비교 결과 <표 2.5>와 같다. 모든 변수들의 등분산 검정결과 재구매 유무에 따라 분산이 동일하지 않았다. t-검정 결과 모든 변수들의 p-값이 0.001미만으로 재구매 유무와 관계가 있는 것으로 판단되었으며, 검정통계량 t의 절대값이 휴면기간, 구매횟수, 평균 구매주기, 평균 구매액 순으로 나타났다.

<표 2.5> 재구매 유무에 따른 변수들의 평균비교

변수	평균		t-값	p-값
	재구매 안함	재구매 함		
구매횟수	2.07	5.39	-94.33	<.0001
총 구매개수	4.67	13.48	-19.62	<.0001
평균 구매개수	2.22	2.35	-17.01	<.0001
총 구매액 (단위:만원)	46.06	134.26	-20.54	<.0001
평균 구매액 (단위:만원)	21.59	23.32	-21.45	<.0001
평균 구매주기	137.63	114.43	31.39	<.0001
휴면기간	459.70	174.91	246.96	<.0001

4. 상관분석

전체 고객들에 대한 변수들의 상관분석 결과 <표 2.6>과 같다. 총 구매갯수와 총 구매액의 상관관계가 가장 높았으며, 평균 구매갯수와 평균 구매액, 구매횟수와 총 구매갯수, 구매횟수와 총 구매액 간의 상관관계가 높게 나타났다. 재구매 유무와 상관관계가 높은 변수로는 휴면기간, 구매횟수였다.

<표 2.6> 상관분석

	구매횟수	총 구매개수	평균 구매개수	총 구매액	평균 구매액	평균 구매주기	휴면기간
총 구매개수	0.61	.	.	.	.	.	.
평균 구매개수	0.04	0.13	.	.	.	.	.
총 구매액	0.60	0.97	0.12	.	.	.	.
평균 구매액	0.05	0.11	0.87	0.13	.	.	.
평균 구매주기	-0.21	-0.05	-0.05	-0.05	-0.06	.	.
휴면기간	-0.28	-0.07	-0.02	-0.07	0.01	-0.13	.
재구매 유무	0.32	0.08	0.03	0.09	0.03	-0.08	-0.33

Ⅲ. 재구매 예측 모형

분석자료에 포함된 변수들을 이용하여 재구매 유무를 예측할 수 있는 모형을 개발하였다. 구매횟수가 1인 경우는 변수 생성과정에서 구매행태에 대한 변수가 결측이 많이 나타나고 있어 좋은 예측모형을 기대할 수 없으므로 총 구매횟수가 2회 이상인 고객들을 대상으로 예측모형을 개발하기로 한다.

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 주어진 분계점에 따라 <표 3.1>과 같이 정오분류표를 작성하여 예측성 평가를 하였다. 본 연구에서는 손건태와 이은혜(2006)와 손건태와 한정임(2004)에서 언급한 Heidke skill score (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, $PCR = (O_1 \mathrm{F}_1 + \mathrm{O}_2 \mathrm{F}_2)/\mathrm{N}^2$ 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> 예측모형에서의 정오분류표

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

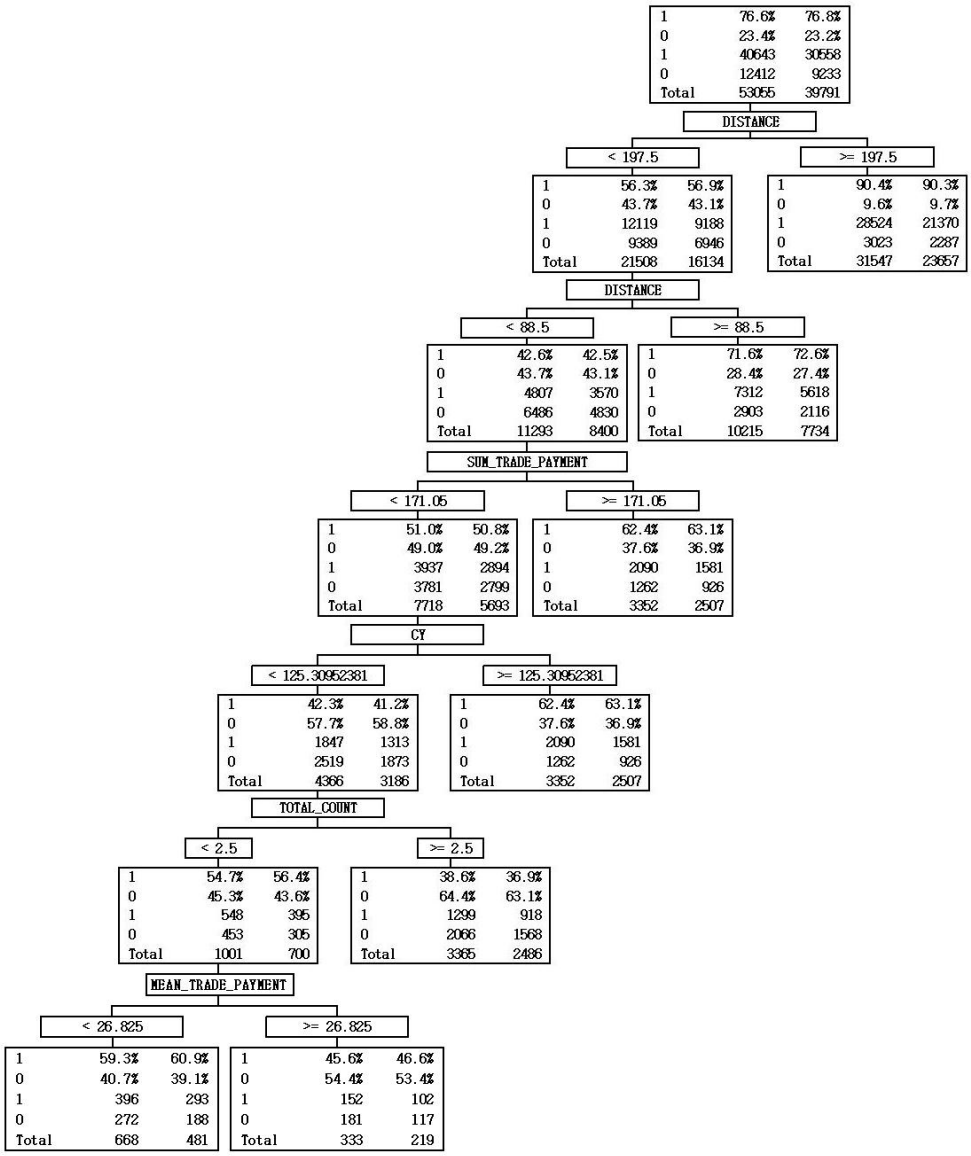
A. 의사결정나무모형

의사결정나무모형 구축을 위해서 분리기준으로 카이제곱 통계량과 엔트로피 지수 및 지니지수를 사용해본 결과 정분류율이 각각 80.22%, 80.26%, 80.22%로서 카이제곱 통계량과 지니지수인 경우 같게 나왔으며 엔트로피지수를 사용할 경우 조금 높게 나타났다. 이를 토대로 엔트로피 지수를 분리기준으로 사용하였다. 분석결과 훈련용 자료, 검증용 자료에서 HSS를 최대로 하는 분계점은 0.29로 동일하였으며, 정분류율은 훈련용 자료에서 약 79.77%, 검증용 자료의 분계점을 토대로 한 평가용 자료에서 약 80.26%였다. 평가용 자료에 대한 정오분류표는 <표 3.2>와 같았으며, 나무구조는 <그림 3.1>과 같다. 가장 중요한 분리변수는 distance(휴면기간)이었으며, 다음으로 sum_trade_payment(총 구매액), cy(평균 구매주기)가 재구매를 예측하는데 중요한 변수임을 알 수 있었다. distance(휴먼기간)가 197.5보다 크거나 같다면 재구매를 할 확률이 90.4%이고, distance(휴먼기간)가 197.5보다 작고 88.5보다 크거나 같다면 재구매를 할 확률이 71.6%가 된다.

<표 3.2> 의사결정나무모형의 정오분류표

		예측		총계
		0	1	
관측값	0	26782 (67.30)	3469 (8.72)	30251 (76.02)
	1	4385 (11.02)	5156 (12.96)	9541 (23.98)
총계		31167 (78.32)	8625 (21.68)	39792 (100.00)

( ) : %



<그림 3.1> 의사결정나무모형

B. 로지스틱회귀모형

로지스틱 회귀모형에서 변수선택방법으로 stepwise방법을 사용하였으며, 7개 설명변수 중 최종 선택된 변수는 cy(평균 구매주기), distance(휴면기간), mean_trade_payment(평균 구매액), total_count(구매횟수), sum_trade_num(총 구매개수)으로 5개 변수이다. 로지스틱회귀모형 결과 <표 3.3>과 같은 회귀계수를 얻었으며, 다음과 같은 로지스틱회귀모형으로 표현된다.

$$Logit(p) = \log(\frac{p}{1-p}) = -0.3373 - 0.0014\,cy - 0.0053\,distance + 0.0053\,mean\_trade\_payment + 0.1042\,total\_count - 0.0037\,sum\_trade\_num$$

<표 3.3> 로지스틱회귀모형에서 회귀계수의 최우추정치

Parameter	DF	Estimate	s.e	Wald-χ^2	p-값
Intercept	1	-0.3373	0.0385	76.89	<.0001
cy	1	-0.0014	0.0001	169.73	<.0001
distance	1	-0.0053	0.0001	4940.69	<.0001
mean_trade_payment	1	0.0053	0.0009	33.72	<.0001
total_count	1	0.1042	0.0036	827.62	<.0001
sum_trade_num	1	-0.0037	0.0005	53.68	<.0001

훈련용 자료와 검증용 자료에서 HSS를 최대로 하는 분계점은 각각 0.37, 0.39였으며, 정분류율은 훈련용 자료에서 79.50%, 검증용 자료의 분계점을 토대로 한 평가용 자료에서 80.63%로 정오분류표는 <표 3.4>와 같다.

<표 3.4> 로지스틱회귀모형의 정오분류표

		예측		총계
		0	1	
관측값	0	26652 (66.98)	3599 (9.04)	30251 (76.02)
	1	4110 (10.33)	5431 (13.65)	9541 (23.98)
총계		30762 (77.31)	9030 (22.69)	39792 (100.00)

( ) : %

C. 신경망 모형

신경망모형의 구조는 하나의 입력층과 하나의 은닉층 그리고 하나의 출력층으로 구성하였으며 활성함수(activation function)은 Hyperbolic Tangent함수를 사용하였다. 은닉층의 뉴런수를 결정하기 위해 여러 뉴런 수에 대해 신경망에 적합신킨 결과는 <표 3.5>와 같으며, SBC(Schwarz Bayesian Criterion)의 값이 가장 작은 2로 하였다. SBC의 경우 추정될 파라미터의 수와 분산의 최우추정값과의 관계를 고려 적절한 차수를 선택하게끔 하는 방법이다. SBC는 $n\times\ln(SSE/n)+p\times\ln(n)$으로 계산되며, 작을수록 좋은 모형이다.

<표 3.5> 뉴런 수에 따른 신경망모형의 비교

뉴런수	Root ASE	Vaild Root ASE	Test Root ASE	Schwarz Bayesian Criterion
1	0.3660	0.3649	0.3642	44902.1298
2	0.3644	0.3635	0.3628	**44644.5278**
3	0.3644	0.3633	0.3629	44728.7325
4	0.3640	0.3628	0.3625	44741.8850
5	0.3635	0.3625	0.3621	44746.4435

최종 신경망모형에 대한 적합결과 각 변수들에 대한 가중치는 <표 3.6>과 같으며, 훈련용자료, 검증용자료에서 HSS를 최대로 하는 분계점은 각각 0.36, 0.39였으며, 훈련용자료에서 정분류율은 약 79.96%, 검증용자료의 분계점을 토대로 한 평가용자료에서의 정분류율은 약 81.09%로 나타났다. 평가용자료에 대한 정오분류표는 <표 3.7>과 같다.

<표 3.6> 신경망 모형에서 각 변수에 대한 가중치

variable	neuron1 (h11)	neuron2 (h12)
cy	-0.5353	0.1935
distance	0.5904	0.6774
mean_trade_num	-0.6573	0.0686
mean_trade_payment	0.2270	-0.0694
sum_trade_num	1.3927	-0.0948
sum_trade_payment	1.2921	0.0568
total_count	-2.1150	-0.0382
bias	-0.6376	1.5420
(to)repurchase	-0.7249	-6.7828
repurchase bias	3.9604	

<표 3.7> 신경망모형의 정오분류표

		예측		총계
		0	1	
관측값	0	26959 (67.75)	3292 (8.27)	30251 (76.02)
	1	4231 (10.63)	5310 (13.34)	9541 (23.98)
총계		31190 (78.38)	8602 (21.62)	39792 (100.00)

( ) : %

D. 모형비교

의사결정나무, 로지스틱회귀모형, 신경망모형의 결과에 대한 비교는 <표 3.8>과 같다. 각 모형의 HSS를 최대로 하는 검증용자료의 분계점은 각각 0.29, 0.39, 0.39였으며, 평가용자료의 정분류율은 신경망모형이 81.09%로 의사결정나무모형 80.26%, 로지스틱회귀모형 80.63% 보다 약간 높게 나타났다. 세모형의 평가용자료로부터 제곱근평균제곱오차(Root ASE)를 비교해 보면 신경망모형이 0.3628로 로지스틱회귀모형 0.3677, 의사결정나무모형 0.3699보다 작으므로 조금 우수한 것으로 나타났다.

<표 3.8> 의사결정나무, 로지스틱회귀, 신경망모형의 비교

모형	분계점*	HSS*	HSS**	정분류율(%)**	Root ASE**
의사결정나무모형	0.29	0.4195	0.4402	80.26	0.3699
로지스틱회귀모형	0.39	0.4363	0.4587	80.63	0.3677
신경망모형	0.39	0.4373	0.4633	81.09	0.3628

*검증용자료, **평가용자료

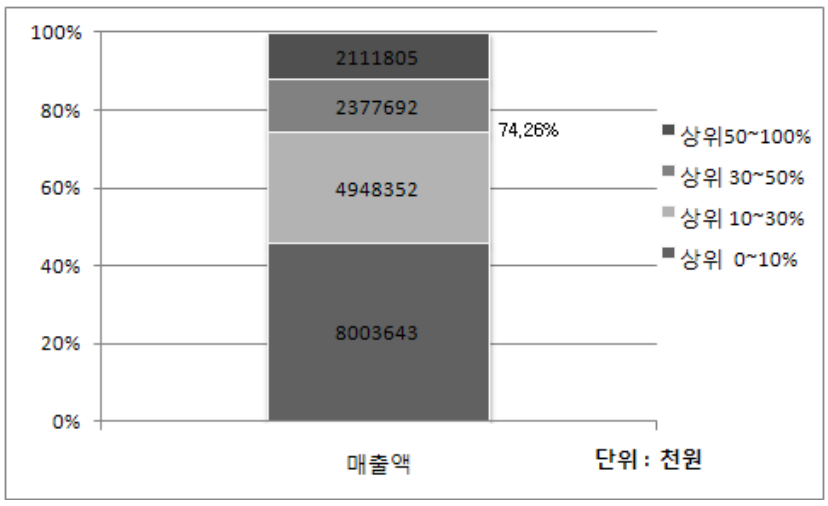
Ⅳ. 고객 분류

고객을 좀 더 효율적으로 관리하기 위하여 고객을 분류하는 경우가 있다. 현재까지 고객의 분류라 하면 성향이 비슷한 고객으로 분류하거나, 고객의 등급제로 분류하는 방법 등이 있다. 고객을 관리하고, 고객을 분류하는 궁극적인 이유는 결국 제품을 팔기 위해서이다. 이 장에서는 제품을 구매할 확률에 따라 고객을 분류하는 경우와 매출기여도가 높은 순으로 고객을 분류하는 경우를 비교하려고 한다.

A. 매출액 순위에 따른 고객 분류

1. 파레토 법칙

기업에서 파레토 법칙을 이용하여 전체 고객을 관리하는 것보다 상위 고객만 잘 관리해도 매출을 유지할 수 있다고 한다. 이 연구 자료에서는 <그림 4.1>과 같이 고객의 상위 30%가 전체 매출의 약 74% 차지하고 있음을 알 수 있었다.



<그림 4.1> 매출액 파레토

2. 매출액 순위에 따른 고객의 등급

전체 매출액의 상위 1%이내 고객을 1등급, 10%이내 고객을 2등급, 30%이내 고객을 3등급, 나머지 고객을 4등급으로 하여 <표 4.1>과 같이 고객을 나누었다.

<표 4.1> 매출액 순위에 따른 고객의 등급

고객의 등급	매출액	N
1등급	상위 0~1%	3006
2등급	상위 1~10%	27155
3등급	상위 10~30%	59997
4등급	상위 30~100%	210351

B. 재구매 확률에 따른 고객 분류

1. 재구매 지수

재구매 예측모형 중 조금 우수하였던 신경망모형에서의 재구매 확률에 100을 곱하여 범위가 0~100 사이에 유지되도록 재구매 지수를 개발하였다. 재구매 지수란 특정 고객이 6개월 이내에 재구매가 일어날 가능성에 대한 지수로 수치가 높을수록 재구매 확률이 높음을 의미한다. 재구매 지수를 이용하여 2005년 1월부터 2007년 12월 까지 3년간 구매 기록이 있는 고객 300509명을 대상으로 고객을 분류를 하였다.

전체 고객 중 3년간 구매횟수가 한 번이였던 고객의 경우 재구매 지수를 산출 할 수 없으므로 전체 고객 중 구매횟수가 2회 이상인 고객 132638명의 재구매 지수의 평균은 23.47이였으며, 재구매 지수의 기초통계량은 <표 4.2>와 같다.

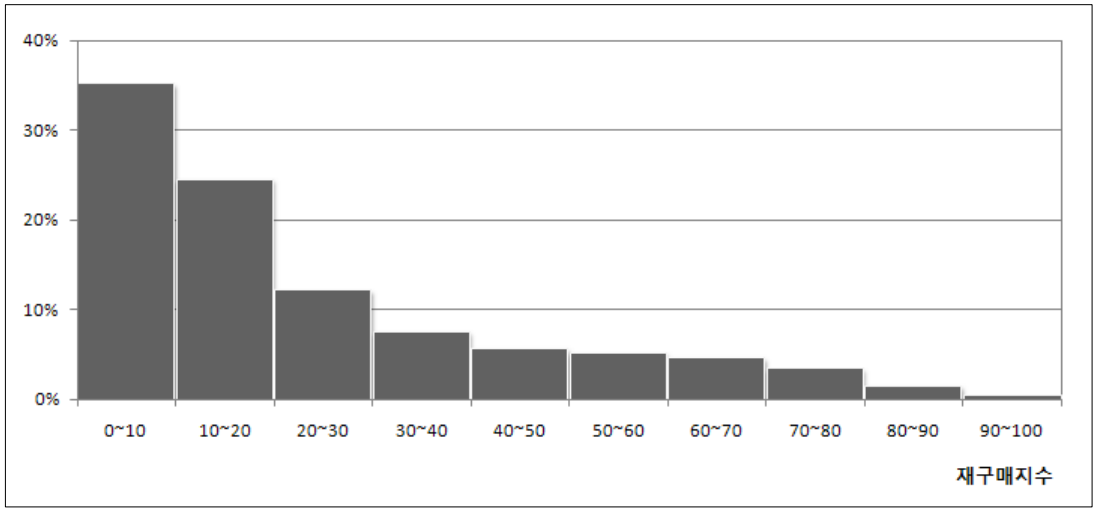
<표 4.2> 재구매 지수의 기초통계량

변수	N	평균값	표준편차	최소값	최대값
구매횟수	132638	23.47	21.62	2.86	100.00

재구매 지수의 분포는 <표 4.3>과 <그림 4.2>와 같다. 지수가 20이하인 고객이 전체의 59.69%로 차지하고 있는 반면 지수가 80이상인 고객은 전체의 1.82%였다.

<표 4.3> 재구매 지수 분포

재구매 지수	고객의 수	백분율(%)
0~10	46636	35.16
10~20	32537	24.53
20~30	16112	12.15
30~40	9952	7.50
40~50	7488	5.65
50~60	6870	5.18
60~70	6129	4.62
70~80	4496	3.39
80~90	1942	1.46
90~100	476	0.36
전체	132638	100.00



<그림 4.2> 재구매 지수 분포

2. 재구매 지수를 이용한 고객 등급

매출액 순위에 따라 분류한 고객 등급별 N수와 동일하게 재구매 지수가 높은 순위로 하여 4개 등급으로 나누었으며, 기준시점 이전 3년간 구매횟수가 한 번이였던 고객의 경우 재구매 지수를 산출할 수 없으므로 4등급으로 분류하여 <표 4.4>와 같이 총 4개 등급으로 나누었다.

<표 4.4> 재구매 지수를 이용한 고객 등급

고객의 등급	지수범위	해석
1등급	78 ~ 100	6개월 안에 재구매 가능성이 78%이상인 고객군
2등급	37 ~ 78	6개월 안에 재구매 가능성이 37~78%인 고객군
3등급	9 ~ 37	6개월 안에 재구매 가능성이 9~37%인 고객군
4등급	0 ~ 9	6개월 안에 재구매 가능성이 0~9%인 고객군 또는 신규 구매한 고객군

3. 고객 등급에 따른 변수들의 평균 비교

고객등급에 따른 변수들의 평균은 <표 4.5>와 같았다. 4등급의 경우 약 80%가 신규고객(구매횟수가 1인 고객)으로 평균구매주기를 구할 때 이를 제외하고 구하였다. 고객등급이 1등급 일수록 구매횟수, 총 구매개수, 평균 구매개수, 총 구매액, 평균 구매액이 많았으며, 평균 구매주기, 휴면기간이 짧음을 알 수 있다.

<표 4.5> 고객등급에 따른 변수들의 평균

등급	구매횟수	총 구매개수	평균 구매개수	총 구매액	평균 구매액	평균 구매주기	휴면 기간
1등급	20.2	25.7	2.5	611.3	62.1	44.1	17.6
2등급	6.7	24.0	2.4	157.6	15.7	91.2	53.1
3등급	3.9	22.0	2.3	87.0	8.8	176.9	239.6
4등급	1.3	21.4	2.2	29.0	2.9	101.4	526.0

C. 고객 등급 비교

전체 매출의 순위에 따른 등급과 재구매 지수를 이용한 고객의 등급을 비교함으로서, 기존의 고객 분류 방식과 통계적 모형을 이용한 고객 분류 방식에 대한 장단점을 간접적으로 알아본다.

1. 고객 등급별 분포 비교

전체 매출의 순위에 따른 등급과 재구매 지수를 이용한 고객의 등급 간 고객의 분포가 차이가 없다면 서로 비교하는 것은 의미가 없을 것이다. 하지만 <표 4.6>과 같이 두 등급간 고객의 분포를 보면, 두 등급 모두 일치하는 고객은 1등급의 경우

51.5%, 2등급은 44.1%, 3등급은 38.9%, 4등급은 85.9%만 일치하는 것을 알 수 있다.

<표 4.6> 고객 등급별 분포 비교

등급		재구매 지수 기준				계
		1	2	3	4	
매출액 기준	1	1547 (51.5)	1010 (33.6)	397 (13.2)	52 (1.7)	3006
	2	1426 (5.3)	11984 (44.1)	11046 (40.7)	2699 (9.9)	27155
	3	31 (0.1)	9713 (16.2)	23346 (38.9)	26907 (44.9)	59997
	4	2 (0.0)	4448 (2.1)	25208 (12.0)	180693 (85.9)	210351
계		3006	27155	59997	210351	300509

( ) : %

2. 기준시점 이후 6개월 동안 재구매 여부 비교

기준시점 이후 향후 6개월 동안 재구매를 여부를 비교한 결과 <표 4.7>과 같다. 매출액 기준으로 고객을 분류한 경우보다 재구매 지수를 기준으로 고객을 분류한 경우 4등급을 제외한 모든 등급에서 재구매율이 더 높았다. 이는 기준시점에서 프로모션을 할 때, 지금까지 매출액이 많았던 고객보다 재구매 지수가 높은 고객을 대상으로 프로모션을 할 경우 더 높은 효율을 얻을 수 있다고 볼 수 있다.

또한, 재구매 지수를 기준으로 고객을 분류한 경우 고객을 대상으로 제품 구매 유도를 위한 프로모션을 계획할 경우 효과적인 투자를 할 수 있다. 예를 들어, 1등급은 재구매할 확률이 78%이상인 등급으로 재구매할 확률이 상당히 높은데, 많은 돈을 투자하여 제품 구매를 유도할 필요가 없으며, 4등급의 경우 제품의 구매가 한 번인 신규고객이 대부분으로 재구매에 대한 정보가 부족하고, 재구매할 확률이 9%

로 이하로 투자대비 효과가 적을 것으로 기대되지만, 고객의 70%정도가 4등급으로 기업의 미래를 위해서 투자가 필요하다. 따라서 4등급의 경우 개별 고객에게 제품 구매유도를 위한 프로모션보다는 고객이탈 방지를 위한, 이탈한 고객을 대체하기 위한 신규고객 유치 프로모션을 하는 것이 효율적이다.

<표 4.7> 고객 등급별 재구매 여부

기준	재구매	1등급	2등급	3등급	4등급	계
매출액	안함	915 (30.4)	16357 (60.2)	48648 (81.1)	193772 (92.1)	259692 (86.4)
	함	2091 (69.6)	10798 (39.8)	11349 (18.9)	16579 (7.9)	40817 (13.6)
재구매 지수	안함	380 (12.6)	12193 (44.9)	48593 (81.0)	198526 (94.4)	259692 (86.4)
	함	2626 (87.4)	14962 (55.1)	11404 (19.0)	11825 (5.6)	40817 (13.6)
계		3006	27155	59997	210351	300509

( ) : %

3. 고객 분류 방법에 대한 장·단점

매출액을 기준으로 고객을 분류한 경우 상위 30%의 고객만 잘 관리해도 전체 매출의 70% 정도를 유지할 수 있었으나, 기준시점에서 광고 등의 프로모션을 통해 향후 제품의 구매를 유도하기 위한 정보가 부족하였다. 재구매 지수를 이용하여 고객을 분류한 경우 제품의 구매를 유도하기 위한 프로모션을 할 때, 효과적인 투자를 계획할 수 있었지만, 기준시점 마다 재구매 예측모형을 개발하여야 되는 단점이 있다.

Ⅴ. 결론 및 향후과제

본 연구는 화장품 구매 자료를 이용하여 과거 3년의 구매행태를 분석하여 이를 바탕으로 향후 6개월 이내의 재구매 여부를 예측하는 모형을 개발하였다. 재구매 여부를 예측하는 모형으로 의사결정나무모형과 로지스틱회귀모형 및 신경망모형을 이용하였다.

모형개발을 위해 사용한 자료는 48개월간의 구매 자료로 고객의 수는 362,432명이고 거래수는 2,224,550건으로 고객 1인당 평균 거래수는 6.14건이다. 이 자료를 이용하여, 각 변수들에 특징을 살펴보고, 재구매 유무에 따른 평균비교와 상관분석을 통하여 재구매에 유의한 영향을 주는 변수를 살펴보았으며, 세 예측모형에서 재구매에 대한 확률을 계산하고 최적의 분계점을 찾기 위해 HSS를 이용하였다. 검증용 자료에서 HSS 값을 최대로 하는 분계점을 기준으로 평가용자료에 적용하여 정오분류표를 작성하였고 예측성 평가를 하였다. 세 모형을 비교한 결과 정분류율은 신경망모형이 약 81.09%로 가장 높고, 다음으로 로지스틱회귀모형이 약 80.63%였으며, 의사결정나무모형이 약 80.26%로 가장 낮게 나타났다. 제곱근평균제곱오차(Root ASE)를 비교해 보면 신경망 모형이 0.3628로 로지스틱회귀모형 0.3677, 의사결정나무모형 0.3699보다 작아 조금 우수한 것으로 나타났다. 세 모형비교에서 차이는 크지 않았지만 신경망모형을 최종모형으로 선택하였다.

재구매 예측 모형을 통해 과거 3년의 구매행태자료를 바탕으로 향후 6개월 이내의 재구매 여부를 예측할 수 있었다. 이를 이용하여 기준시점에서 향후 6개월 동안의 판매를 예측할 수 있을 것이다. 또한 재구매 예측모형을 통해 재구매 확률을 구할 수 있었다. 이를 이용하여 기준시점에서 향후 6개월 이내 재구매할 확률이 높은 고객과 낮은 고객에 대한 차별화된 마케팅 전략을 수립 할 수 있을 것이다.

고객을 분류하는데 있어 매출액을 기준으로 고객을 살펴본 결과 상위 30%의 고객의 전체 매출의 70%이상을 차지하고 있었다. 이는 상위 30%의 고객만 잘 관리해도 전체 매출의 70% 정도를 유지 할 수 있음을 알 수 있다. 재구매 예측모형으로

가장 우수하였던 신경망모형을 통해 향후 6개월 이내 재구매 확률을 구할 수 있었다. 이를 이용하여 재구매 지수를 산출하여 고객을 분류하였다. 1등급은 재구매율할 확률이 78%이상인 등급이었으며, 4등급의 경우 재구매할 확률이 9%로 이하로 상당히 낮았다. 향후 6개월간 마케팅 계획을 수립할 경우 등급별로 투자를 해야 할 그룹과 투자를 하지 않아야 할 그룹을 생각해 볼 수 있을 것이다.

본 논문에서는 기준시점에서 향후 재구매 확률을 이용하여 마케팅에 도움이 될 수 있는 방법을 제시하였다. 이는 기준시점이 달라짐에 따라 재구매 확률을 다시 예측해야 되는 번거로움이 있다. 또한 보다 정교한 재구매 예측모형을 개발하기 위해 보다 많은 정보와 연구가 필요하다고 생각한다.

기존 고객들의 자료를 바탕으로 분석을 함으로 인해, 신규고객에 대한 정보가 부족하여 신규고객을 확보 할 수 있는 방법은 제시할 수 없었다. 기존고객에 대한 관리도 중요하지만 신규고객의 확보도 중요한 문제로 신규고객에 대한 연구도 필요하다.

참고문헌

[1] 강현철, 한상태, 신혜림 (2003). 데이터마이닝 기법을 이용한 고객생애가치 측정 모형 개발, Journal of the Korean Data Analysis Society, Vol. 5, No. 4, 927-936.

[2] 김이태 (2005). CRM 고객관계관리, 대경출판사.

[3] 김순귀 정동빈 박영술 (2003). spss를 활용한 로지스틱 회귀모형의 이해와 응용, SPSS 아카데미.

[4] 백신정 (2004). 데이터마이닝을 통한 고객관리데이터의 분석, 석사학위논문, 고려대학교.

[5] 손건태 (2006). Binary Forecast of Heavy Snow using Statistical Models, The Korean Communications in Statistics, Vol. 13, No. 2, 369-378.

[6] 조대현, 김병수, 석경하, 이종언, 김종성, 김선화 (2009). 화장품구매 자료를 통한 고객 구매행태 분석, Journal of the Korean Data & Information Science Society, 20, 615-627.

[7] 조대현, 김병수, 석경하, 이종언, 오동관, 김종성, 김선화 (2008). 마이크로마케팅 방법론 개발 : 고객 충성도 예측모형, Journal of the Korean Data Analysis Society, Vol. 10, No. 4, 2075-2087.

[8] 최종후, 한상태, 강현철, 김은석(1998). 데이터마이닝의 의사결정나무분석, 고려정보산업.

[9] 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.

[10] Heidke, P. (1926). Berechnung des Erfolges und der Gute der Windstar kevorhersagen im Sturmwarnungsdienst, Geografiska Annaler, 8, 301-349.

[11] Judy Strauss and Raymond Forst (2001). E-Marketing, Prentice Hall.

[12] 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.

[13] Neter, J., Kutner M. H., Nachtsheim C. J., Wasserman, W. (1996). Applied Linear Regression Models, Third Edition, Irwin, Chicago.