經營論集, 第 37 卷 2 · 3號 (2003年 9月)

고객 정보의 종류와 양이 구매모형 예측력에 미치는 영향에 관한 연구

오 윤 경*
김 지 경**
김 상 훈***

이 글에서는 일대일 마케팅 전략에 있어 핵심적으로 다루어지는 구매여부, 구매시점, 구매금액 등 세 가지 목적변수를 예측하기 위한 예측 모형들의 예측력이 고객 정보의 종류와 수에 따라 어떻게 변하는지 고찰하였다. 그 결과 인구통계정보와 구매이력정보 모두 모형의 예측력을 증가시켰지만, 후자의 경우에 더욱 두드러지게 증가시켰다. 또 구매이력 정보의 수가 증가할수록 모형의 예측력이 증가하는 것으로 나타났다. 단 구매이력정보의 수가 증가할수록 모형의 한계개선 정도는 감소하였다.

모형의 예측력 증가로 인한 타겟 마케팅 효과의 증대는 기업의 수익과 이익으로 직접 환산될 수 있다. 이 글의 후반부에서는 구매이력 정보의 획득 비용이 증가함에 따라 기업의 이익이 어떻게 변하는지 시각적으로 제시하였다.

이 같은 연구결과는 고객 데이터를 수집하고 활용하는 기업전략이 수익성 증대에 기여할 수 있다는 점을 실증적으로 증명하고 있다. 나아가 기업의 일대일 마케팅 효과를 극대화하기 위해 고객 데이터 획득 방법과 수준을 전략적으로 결정할 때 실질적인 기여를 할 수 있을 것으로 기대한다.

I. 서 론

오늘날 기업은 디지털 테크놀로지를 이용하여 방대한 양의 고객정보를 수집하고

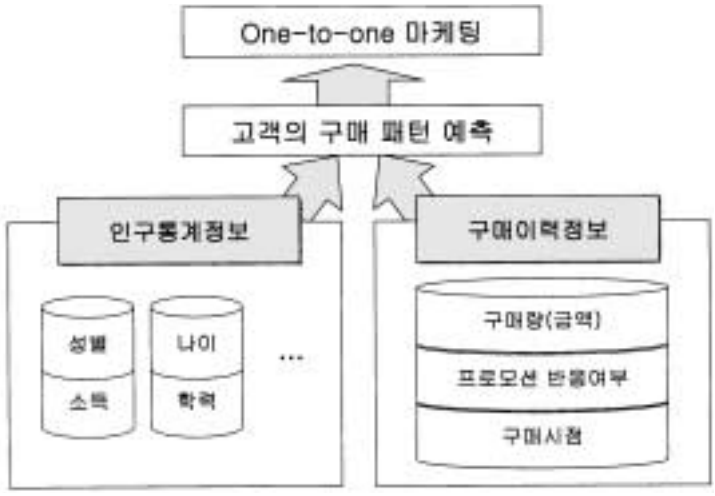
*퍼듀대학(Purdue University) 박사과정 : oh13@purdue.edu
**서울대학교 대학원 경영대학 석사 : jeanne@hanafos.com
***서울대학교 경영대학 조교수 : profkim@snu.ac.kr

그것을 분석하여 효과적인 대(對)고객 전략을 세울 수 있게 되었다. 인터넷으로 통칭되는 네트워크의 발달은 고객 접점 데이터를 빠르고 저렴하며 대량으로 수집할 수 있게 하였고, 컴퓨팅 시스템의 발달은 그렇게 입수된 방대한 데이터를 비용 효율적이고 다각적으로 분석할 수 있게 하였다. 나아가서 생산에서부터 판매, 사후관리에 이르기까지 고객과 긴밀한 관계를 유지하는 전략이 가능해졌다.

이러한 디지털 기술의 마케팅적 활용에 있어 대명사와 같이 사용되는 개념이 바로 고객 관계관리(CRM: customer relationship management)이다. CRM이란 고객관리에 있어 필수적인 요소들, 즉 기술 인프라, 시스템 기능, 사업전략, 영업프로세스, 조직의 경영능력, 고객과 시장에 관련된 영업정보 등을 고객 중심으로 정리, 통합하여 대(對)고객활동(customer interaction)을 개선함으로써 고객과의 장기적인 관계를 구축하고 기업의 경영성과를 개선하기 위한 새로운 경영 방식이다〔이유재 외, 2001〕. 최근 CRM의 가장 일반적인 활용방식은 기존 고객의 프로파일을 바탕으로 가망 고객을 예측하는 고객획득(customer acquisition), 이미 자사 제품을 구입한 고객들을 대상으로 재구매를 유도하는 고객유지(customer retention), 고객 데이터 베이스를 바탕으로 자사의 타 제품을 추가/교차판매 하는 활동(up-selling and cross-selling) 등 세 가지이다. 실제 기업들은 CRM의 실무적 활용으로써 일대일 마케팅(one-to-one marketing)을 수행하고 있는데, 이것은 고객 정보를 바탕으로 개개인의 구매 성향을 파악한 후 차별화 된 마케팅 프로모션을 통하여 매출을 극대화하는 것이다.

이와 관련하여 가장 기본이 되는 것이 고객의 구매행동 예측 모형이라 할 수 있다. 이것은 고객이 언제 구매하고, 얼마나 구매하며, 마케팅 노력에 대한 반응율이 어느 정도인가를 정량적으로 평가하는 모형이다. 이러한 모형 수립의 기반이 되는 고객 데이터는 크게 인구통계학적 정보와 구매이력정보로 나눌 수 있는데, 특히 구매이력정보는 고객의 미래 구매행동을 예측하는데 보다 실질적이고 직접적이라는 측면에서 그 가치가 높다고 할 수 있다〔Rossi et al., 1996〕. 이러한 고객 데이터는 기업의 중요 자산으로까지 인식되어 기업의 인수 합병 시 기업 가치를 높이는 데 기여하고 있다. 고객 데이터의 중요성을 인지하고 있는 많은 기업들은 광범위한 고객 데이터 수집을 위해 데이터 웨어하우스(Data warehouse)를 구축하고 있으며, 양질의 데이터를 획득하기 위하여 데이터 수집 기관으로부터 고객 데이터를 구입하는 경우도 늘고 있다.

이러한 상황에서 고객 정보의 가치를 실증적으로 평가하고, 특히 기업에 축적되는



〔그림 1〕 고객 정보 유형과 일대일 마케팅에의 활용

구매이력정보의 가치를 정확하게 평가하는 것은 중요한 의미를 가지고 있다 하겠다.

II. 기존 문헌 연구

1. 고객 정보의 가치에 관한 연구

1970년대에 들어서 과잉 공급으로 인한 전지구적 경제 침체기를 거치며 기업들이 기존에 가지고 있던 생산 중심적 사고가 판매 중심적 사고로 급선회 하게 되었다. 마케팅의 중요성이 강조되는 시기를 거쳐 최종적으로는 고객과의 장기적이고 우호적인 관계, 더 나아가 고객 그 자체가 가치 평가의 대상이 되는 현재에 이르렀다. 즉 고객과의 관계는 유동자산이나 고정자산과 같은 물질적 자산과 마찬가지로 기업의 중요한 자산을 이룬다는 인식이 보편성을 얻고 있는 것이다〔이유재 외, 2001〕.

고객 자산(Customer Equity)은 재무적인 자산으로서 가치 측정, 관리의 대상이 된다. 최근에는 고객 자산 관리 비용 절감 혹은 수익의 극대화의 두 가지 접근방식을 통한 마케팅 투자대비 수익성(ROI: Return on Investment)을 평가하는 것이 중요한 자

산 평가 절차의 일부로 자리잡았다. Blattberg et al.(2001)은 이 같은 평가방식의 유효성을 다음과 같이 정리하였다. 첫째, 신규고객획득(acquisition), 기존고객유지(retention), 교차판매(add -on selling)의 주요 마케팅 활동에 투자할 경우에 증가할 것으로 기대되는 고객 자산가치를 산출할 수 있다. 둘째, 고객과 동적으로 관계를 맺는 과정에서 마케팅 투자비용의 수준을 결정할 수 있다. 셋째, 개별 고객 생애 주기(customer life cycle)를 고려한 수익성 극대화라는 전략적 시각에서 신규고객획득, 기존고객유지, 교차판매를 구상할 수 있다. 넷째, 자사 제품을 전반적으로 많이 이용하는 "완전 고객(whole customer)"을 분류하고 규명할 수 있다. 다섯째, 새로운 고객을 획득하고 관계를 강화하는 데 타 고객들과의 상호작용을 활용할 수 있다.

고객 자산을 관리하기 위해서 기업은 고객의 정보로 이루어진 데이터베이스를 먼저 구축해야 한다. 데이터베이스에 대한 통찰 없이 고객의 향후 소비 행동을 예측하기는 어렵기 때문이다. 기업은 고객의 데이터베이스를 통해 자사의 가장 중요한 고객이 누구인지를 알아낼 수 있으며 그들의 과거 구매 내역을 살펴봄으로써 가장 반응률을 높이는 커뮤니케이션 방법을 찾아낼 수 있다. 나아가 과거의 구매 내역에 인구통계적 정보나 심리적인 자료를 추가하여 분석하면 미래의 구매에 대한 예측도 가능할 것이다〔Stone, 1996〕.

수익성이 가장 높은 고객을 찾기 위해 고객 정보를 이용하는 기법으로 가장 널리 사용되어온 방법은 RFM 분석기법이다. RFM 분석기법은 고객의 구매 행동을 구매시기/빈도/구매금액(Recency/Frequency/Monetary)의 세 가지 척도에 의거하여 각각의 고객들의 기록에 따라 수익성을 평가하는 방법이다. RFM 기법은 고객의 미래 잠재가능성을 예측하게 해주며 앞으로의 판매촉진 비용을 데이터베이스 상의 각각의 고객들로부터 얻을 수 있는 잠재이익과 비교할 수 있도록 해준다. 산업이나 기업의 상황에 따라 RFM의 세 가지 변수들이 구매에 미치는 영향을 측정해서 가중치를 부여하여 계산하기도 한다.

RFM 분석기법을 통계적인 용어로 표현한다면 주효과(main effect)와 모든 상호작용효과(interaction)를 고려한 삼원분산분석(three-way ANOVA) 이라 할 수 있다. RFM 분석기법을 삼원분산분석으로 해석하면 기존의 RFM 모형을 확장하여 적용할 수 있다는 장점이 있다. 많은 실무자와 연구자들은 고객의 반응확률을 예측하는데 있어 RFM뿐만 아니라 제품의 선택 및 인구통계적 변수 등 다른 변수들을 고려해야 한

다고 주장하고 있다〔McEachern, 1998; Miller, 1998; Schmid and Weber, 1998; Girard, 2000〕.

한편, Rossi et al.(1996)은 타겟마케팅을 목적으로 하여 사용되는 인구통계학적 정보와 구매이력정보에 대한 평가를 시도하였다. 이를 위해 개별 가구 수준에서의 관측가능하거나 관측 불가능한 상호이질성(observable and unobservable heterogeneity)을 포괄할 수 있는 새로운 랜덤 계수 모델(flexible random coefficient model)을 개발하였다. 이 모형을 타겟화된 쿠폰을 발급하는데 사용했을 경우를 상정하고, 다양한 정보의 조합을 독립변수로 사용하면서 계층적 베이지안 모수 추정방법을 이용하여 계수를 추정하였다. 연구 결과, 개별 가구의 구매이력 정보까지 사용하여 쿠폰을 발송하였을 경우, 무작위로 쿠폰을 발행했을 때에 비해 2.5배의 수익의 증가를 가져온다는 결과를 실증적으로 도출하였다.

2. 일대일 마케팅 효과에 관한 연구

CRM에 근거한 일대일 고객관리 전략은 정보기술(IT)의 비약적인 발전을 등에 업고 이제는 '선택'이 아닌 '필수'로 자리 잡고 있다. 물건을 얼마나 싸게 잘 만들 수 있는가도 중요하지만, CRM을 제대로 활용해 상품가치와 실제수요를 극대화시킬 수 있는가의 여부가 기업들의 운명을 결정하는 시대가 도래한 것이다.

CRM이란 고객이 원하는 것이 무엇인지 발견하고, 원하는 것을 원하는 시간에 원하는 방법으로 제공하는 새로운 개념이다. 90년대 초반 다국적기업인 IBM이 고안한 것으로 알려졌으며, 국내에는 98년 초 처음 소개돼 지난해부터 본격적으로 회사 내에 전담팀들이 구성되기 시작했다. CRM이 도입되면서 기업들은 데이터를 통합하여 관리하고 부서간 정보 공유를 함으로써 생산성 향상을 가져오게 되었다. 영업사원이 제품설명서와 고객에 관한 정보를 신속하게 파악할 수 있게 되어 제품 판매 과정이 손쉬워지는 것은 물론, 마케팅 담당자가 고객의 니즈와 구매 이력을 자세하게 파악하여 목표 판매량을 보다 정확하게 설정할 수 있게 되었기 때문이다.

그러나 시설 투자에 너무 과대한 비용을 소모하거나, 산업의 특성을 고려하지 않은 채 무조건적으로 도입하는 경우에는 무리한 비용부담을 안게 되거나 기존 조직 문화와의 마찰이 발생하여 실패하는 사례도 적지 않게 보고되고 있다.

CRM 개념보다 역사가 오래된 것으로, 인터넷과 같은 기술적인 혁신이 이루어지기 이전에 기업들이 주로 사용해온 마케팅 방법인 다이렉트 마케팅(Direct Marketing)이 있다. 다이렉트 마케팅은 제품이나 서비스의 교환에 영향을 미치는 과정에서 직접 판매, 직접 우편, 텔레마케팅, 직접 행동 광고, 카탈로그 판매, 케이블 TV 판매 등과 같은 매체 중 하나 또는 하나 이상을 이용하여 예상고객 또는 현재 고객으로부터 전화, 우편, 개인적 방문과 같은 직접 반응을 이끌어 내기 위한 노력을 지휘하는 총체적인 활동을 일컫는다.

다이렉트 마케팅은 미국의 우편 서비스의 발달과 함께 도시와 외곽지역에 거주하는 사람들에게 카탈로그를 보내게 됨에 따라 발전하게 되었으며 최근에는 컴퓨터, 통신 기술과 신용카드의 발전과 더불어 일반적인 마케팅 기법으로 인식되게 되었다. 이러한 발전과 함께 다이렉트 메일 산업의 주요한 난제 중의 하나는 타겟팅의 오류를 최소화 하여 소비자가 원하지 않는 메일을 받게 됨으로 인한 낭비를 최소화 하고 반응률을 높이는 것이다.

다이렉트 마케팅(혹은 메일링)에 대한 연구는 관련 산업의 발전과 함께 90년대 초 이후 많은 이슈들을 다루어왔다(Bawa and Shoemaker, 1987; Bult, 1993; Wedel et al., 1993). Blattberg와 Deighton(1991)은 처음으로 상호작용적인(interactive) 마케팅의 효과에 관하여 논하였으며, DeSarbo와 Ramaswamy(1994)는 새로운 알고리즘을 통해 고객 반응에 있어서의 이질성을 측정하였다. Schmittlein과 Peterson(1994)는 기업간(B-to-B) 마케팅 모델을 산업 데이터베이스에 적용했으며 언제 고객을 리스트에서 제외시켜야 하는 지에 관해 예측하였다. 그들은 모델에서 기대 이익이 양수일 때만 고객을 리스트에 추가하도록 제안하였다. 또한 Bitran과 Mondschein(1996)은 회사의 이익 극대화 지점을 찾는 휴리스틱 방법을 제시하였다.

한편, 이어지는 연구에서 Bult and Wansbeek(1995)는 다이렉트 마케팅을 위한 메일링 리스트를 작성하는 과정에서 분석적인 틀을 도입하였다. 즉, 한계비용과 한계수익이 일치하는 최적 지점을 찾음으로써, 수익 극대화 관점에서, 타겟 메일을 보낼 것인가의 여부를 결정하는 것이다.

3. 구매 예측 모형에 관한 기존의 연구

관측된 고객 데이터를 기반으로 미래의 소비자 행동을 이해하고자 하는 노력은 다양한 관점에서 시도되어 왔다. 소비자의 구매 의도를 예측하는 것은 해당 기업체뿐만 아니라 관련 산업에 있어서도 매우 중요하기 때문이다. Schimittlen et al.(1987)은 고객의 과거 구매 성향을 구매 빈도와 시점, 구매 금액의 세 가지 차원에서 분석하여 해당 시점에 얼마나 많은 고객이 얼마 만큼을 구입할 것인지에 대한 정보를 추론하고자 하였다.

그 이후에도 학계에서는 미래 소비자의 행동을 예측하기 위한 모델을 수립하려는 시도를 다각도로 전개하였는데 여기서는 크게 소비자의 구매 여부와 구매 시점의 두 가지 측면에서 살펴보았다.

1) 구매 여부 예측에 관한 연구

소비자의 구매 여부를 예측하는 경우 종속변수는 구매/비구매의 이산적인 성격을 가지고 있다. 종속변수가 이산적일 경우 전통적인 회귀모형을 사용하는 것은 적절하지 않기 때문에 소비자의 구매 의도를 효과적으로 측정하기 위해 회귀분석 모델이나 로짓(logit), 프로빗(probit) 모형이 많이 사용되어 왔다. 최근에는 산업 공학에서 개발된 인공신경망(neural network) 방법과 같은 귀납적 학습 방법이 도입되어 비선형성의 문제를 해결하며 보다 정확한 소비자 구매 예측에 이용되고 있다. 국내에서는 한상만(2000)이 내구재의 구매의도 예측하는 데 있어 인공신경망을 이용한 방법이 로짓모형을 이용한 방법에 비해 예측력이 우수하다는 사실을 실증한 바 있다.

2) 구매 시점 예측에 관한 연구

사건의 발생이 지속기간에 따라 달라지는 경우를 분석할 때 주로 사용되는 기법이 위험모형(hazard model)이다. 이 기법은 어떤 시점에서 특정 사건이 일어날 가능성은 지속기간에 영향을 받는다는 것을 전제로 하고 있다. 예를 들어 한 영업사원이 직장을 그만두게 될 확률은 처음 2년 동안은 증가하다가 그 이후에는 감소하게 된다. 이 모형은 마케팅에서 구매 간격에 관한 연구나〔Schmittlein and Morrison 1983b; Dunn,

Reader and Wrigley, 1983〕 광고 노출 효과에 관한 연구〔Green. 1982〕, 신제품 수용 확률에 관한 연구(adoption rate)〔Bass, 1969; Schmittlein and Mahajan, 1982〕 등에 널리 이용되어 왔다.

III. 연구 목적과 모형

1. 연구의 목적

본 연구에서는 구매여부, 구매시기, 구매금액을 예측하기 위한 세 가지의 모형을 측정집단 데이터에 기반하여 제시하고, 검증집단의 데이터를 사용하여 각 모형의 예측력을 평가한다. 이때 각 모형의 예측력이 데이터의 종류와 양에 따라 어떻게 변하는지 알아보고자 한다. 이 글을 통하여 규명하고자 하는 문제들은 구체적으로 아래와 같다.

첫째, 고객에 대한 정보를 바탕으로 한 구매예측 모형들이 그렇지 않은 모형들에 비하여 예측력이 유의하게 높아지는가를 알아본다.

둘째, 인구통계 정보만을 사용한 모형보다 구매 이력 정보를 추가하였을 때 모형의 예측력이 높아지는가를 알아본다.

셋째, 구매이력정보가 많아질수록 모형의 예측력이 증가하는지 알아본다.

넷째, 구매이력정보의 증가에 따른 모형 예측력의 증가가 체감하는지 알아본다.

2. 데이터의 수집

본 연구에 사용된 자료는 미국의 다이렉트 마케팅 연구기관인 DMEF(Direct Marketing Educational Foundation)에서 연구용으로 수집한 패널데이터로 어느 비영리기관에서 86년 10월부터 95년 12월까지 약 100만 명의 사람들에게 기부 권유메일을 보내고(solicitation) 이에 응답하여 기부한 이력(contribution)을 각각 10회에 걸쳐 추적한 데이터이다. 이 데이터에는 권유 메일의 발송 시점과 그것을 받은 사람들이 언제, 얼마나 기부했는가 등의 수치가 나타나 있다. 또 우편번호를 기반으로 가계데이

〈표 1〉 데이터의 구성과 적용

유 형	변수 레이블 (variable label)	실제의미	마케팅적 적용
인구통계정보 (demographic info.)	ID	기부자(가계) ID	고객 ID
	gender	성 별	
	income	소 득	
	edu	학 력	
구매이력정보 (purchase history info.)	sldate 1-10	기부 권유 메일 발송일 (solicitation date)	프로모션 시점
	cndate 1-10	기부 시점 (contribution date)	구매 시점
	resdol 1-10	기부 금액 (contribution dollar)	구매 금액

터와 연계하여, 기부자(가계)의 소득, 인종 구성, 학력 등의 인구통계학적 변수도 함께 추적할 수 있다. 총 16,174개의 대상 샘플을 측정집단(calibration set)과 검증집단(validation set)으로 분리하고 측정집단을 통해 모형의 계수를 산출한 후, 검증집단으로 모형의 예측력을 평가하여 비교하였다.

본 연구에서는 이 데이터를 마케팅 상황에 맞게 변용하여, 기부 권유메일을 보낸 시점을 타겟 메일 발송 시점으로, 기부 날짜와 기부 금액을 구매 날짜와 구매 금액으로 재정의하여 분석에 사용하기로 한다.

모형에 사용되는 데이터는 크게 인구통계정보와 구매 이력 정보로 분류할 수 있다. 구매이력정보의 경우 기준 시점이 되는 t_0기의 한 단계 이전인 t_0-1기의 관측치들로부터 가장 오래된 관측치인 t_0-10 시점에 해당하는 구매이력정보를 사용하였다. 정리하면 〈표 1〉과 같다.

3. 모형 설계와 가설

한 기업에서 일대일 마케팅을 수행하고자 하여 고객의 미래 구매 패턴을 예측하고

자 할 때, 가장 관심이 되는 목적변수는 회사의 매출 구조에 따라 상이할 것이다. 예를 들어 자동차와 같이 소비자의 단 한 번의 구매가 회사의 수익성에 큰 영향을 미치는 경우에는 구매 여부가 주된 이슈일 것이며 화장품과 같이 반복적인 구매가 일어나는 제품의 경우 고객의 구매 주기를 파악하여 시기에 맞는 프로모션을 하는 것이 중요한 이슈일 것이다.

이러한 문제 의식을 바탕으로 본 연구에서는 구매여부, 구매시점, 구매금액의 세 가지의 변수들을 마케팅의 중요한 목적변수로 상정하고 각각을 예측하는 데 있어 독립변수로 사용되는 고객의 정보 유형이 모형의 예측력에 어떤 영향을 미치는가를 규명하고자 하였다.

이를 위해 문헌 연구에 근거하여 각각의 목적변수를 가장 잘 설명해줄 수 있는 모형으로 각각 판별분석, 위험함수분석, 선형회귀분석을 선정한 후, 각 모형의 예측 성과를 판단해줄 수 있는 지표를 선택하여 고객의 정보집합의 변화에 따라 모형의 예측력이 어떻게 달라지는 지를 살펴보고자 하였다. 각각을 검증하기 위해 사용된 모형과 예측력 측정 지표는 〈표 2〉와 같다.

한편, 각 모델에 공통적으로 사용되는 독립변수들을 다음과 같은 세 종류로 구분하였다. 첫째는, 고객에 관해 아무 정보가 없을 때이고 둘째는 성별, 나이와 같은 인구

〈표 2〉 예측모델과 분석방법

예측모델	분석방법	예측력 측정지표
구매여부	판별분석	오분류율
구매시점	위험모형	RMSE
구매금액	다중회귀분석	RMSE

〈표3〉 정보 집합의 구성

종 류	설 명
기본(Base)	모집단에서 각 변수들의 분포에 관한 정보: 개별 고객에 관한 특별한 정보가 없음
인구통계(Demo)	개별 고객의 인구 통계 정보
구매 이력 1-10 (PH1-PH10)	인구통계+개별 고객의 구매이력 정보: 개별 고객의 프로모션에 대한 반응 기간과 구매 금액에 관한 정보(최근 정보로부터 10개)

통계정보 만을 사용했을 경우이다. 마지막은 구매이력정보들 — 과거 구매시점과 구매금액 — 을 독립변수로 추가로 사용한 경우로서 최근의 데이터부터 순서대로 10개의 구매 정보를 이용하였다(〈표 3〉 참조).

Ⅳ. 실증분석 결과

1. 구매 여부 예측 모형-판별분석

아래에서는 구매 프로모션 목적으로 보낸 타겟 메일에 대한 고객의 과거 반응 기록을 분석하여 구매 여부를 예측하는 모형을 수립할 때, 모형의 예측력이 얼마나 개선되는가를 검증하고자 한다. 구체적으로는 모형에 독립변수로 사용된 정보 집합이 달라짐에 따라 모형의 예측력이 개선되는 데 있어 얼마나 영향을 미치는 가를 살펴보기 위함이다.

이를 위해 본 연구에서 도입한 분석 기법은 판별분석으로 이것은 집단들 사이의 특성 차이를 극대화 시켜줄 수 있는 변수들의 선형결합을 도출하여 관측 자료를 분류하는 기법이다. 이를 이용하여 기준 시점에서 구매한 집단과 비구매 집단을 분류하는 판별식을 도출하면, 소비자의 구매에 영향을 주는 변수들을 찾아낼 수 있기 때문에 구매 여부 예측에 효과적으로 활용할 수 있다(〈표 4〉 참조). 판별분석 결과는 〈표 5〉와 같다.

추정된 판별 계수들에 대한 해석을 용이하게 하기 위해 집단간 판별계수의 차이를 우측에 함께 명시하였다. 이를 살펴보면 dur 1-dur 10 변수들의 구매집단과 비구매집

〈표 4〉 판별분석 모형요약

변 수		판별분석
종속변수(Y)	response	t_0기에서 구매 여부
독립변수(X)	gender/income/edu resdol 1-resdol 10 dur 1-dur 10	인구통계변수 - 성별, 소득, 학력 t_0-1기에서 t_0-10기의 구매금액 t_0-1기에서 t_0-10기의 프로모션 반응간격

〈표 5〉 판별함수의 추정 계수

변수명		y = 1(구매)	y = 0(비구매)	집단간 차이 (Difference)	P-Value	
Intercept		-34.9065	-36.6580	1.7514		
구매이력정보	dur 1	0.0612	0.0692	-0.0081	<.0001	***
	dur 2	0.1371	0.1428	-0.0057	<.0001	***
	dur 3	0.0577	0.0607	-0.0030	<.0001	***
	dur 4	0.0255	0.0313	-0.0058	<.0001	***
	dur 5	0.0363	0.0370	-0.0007	<.0001	***
	dur 6	0.0374	0.0391	-0.0017	<.0001	***
	dur 7	0.0317	0.0340	-0.0023	<.0001	***
	dur 8	0.0860	0.0865	-0.0005	0.8800	
	dur 9	0.0273	0.0283	-0.0009	0.0049	***
	dur 10	0.0257	0.0261	-0.0004	0.0872	*
	resdol 1	0.5961	0.5158	0.0803	<.0001	***
	resdol 2	1.8527	1.7798	0.0730	<.0001	***
	resdol 3	1.1681	1.0714	0.0967	<.0001	***
	resdol 4	0.2900	0.2804	0.0095	<.0001	***
	resdol 5	0.1343	0.1297	0.0046	<.0001	***
	resdol 6	0.2787	0.2747	0.0040	0.0019	***
	resdol 7	0.1711	0.1678	0.0033	0.1269	
	resdol 8	0.5120	0.5233	-0.0113	0.4100	
	resdol 9	0.0840	0.0783	0.0057	0.9813	
	resdol 10	0.0451	0.0624	-0.0173	<.0001	***
인구통계정보	gender	2.9950	3.0006	-0.0056	0.9822	
	income	0.1770	0.1774	-0.0004	0.2719	
	edu	-0.1337	-0.1349	0.0012	0.3008	

단, ***는 유의수준 0.01 에서 유의, **는 0.05에서 유의, *는 0.1에서 유의

단 간의 차이는 모두 음(-)의 값을 나타내므로 타겟 광고 메일에 반응하기 까지 걸리는 시간이 짧을수록 구매 집단으로 분류할 수 있을 것이다. 또한 resdol 8과 resdol 10을 제외한 과거 구매금액에 대한 변수들의 집단간 계수의 차이는 모두 양(+)을 값을 나타내므로 과거 구매액이 많을수록 구매를 하는 집단 쪽으로 분류된다고 할 수 있으며 이것은 우리의 직관과도 잘 일치하는 결과라고 할 수 있다.

이렇게 도출된 판별함수를 분류에 이용하기 위해 여기서는 두 집단의 사전확률 p_1, p_2을 고려한 일반화된 거리에 의한 판별함수 방법을 사용하였다. 공분산의 동일성 검정에 의해 두 집단의 분산이 다르다는 것이 확인되었으므로 한 개체 x로부터 각 모집단의 중심까지의 거리를 나타내는 일반화된 거리제곱(generalized squared distance)은 다음과 같이 정의된다.

$$D_j^2(x) = (x - \overline{x_j})' S_p^{-1}(x - \overline{x_j}) + \ln|S_j| - 2\ln(P_j), j = 1, 2 \quad (1.4)$$

이것을 이용하여 한 개체 x가 j번째 모집단에 속할 사후확률(posterior probability)을 구하면 (1.5)의 수식과 같으며 이 값이 최대가 되는 모집단으로 새로운 샘플을 분류한다.

$$P(j|x) = \frac{\exp(-0.5D_j^2(x))}{\sum_{k=1}^{2}\exp(-0.5D_k^2(x))} \quad (1.5)$$

한편, 판별분석을 수행하여 도출된 판별함수가 관측치의 좋은 분류기준이 되기 위해서는 가능한 잘못 분류될 확률 또는 잘못 분류되었을 경우 발생할 수 있는 비용이 최대한 작아야 한다. 본 연구에서는 (1.5)의 식을 이용하여 측정집단과 검증집단을 분류한 후, 판별분석의 적합도(goodness of fit)을 평가하는 여러 지표 중에서 식 (1.6)과 같이 정의되는 총 오분류확률(total probability of misclassification: TPM)을 지표로 삼아 독립변수로 사용되는 정보집합이 달라짐에 따라 모델의 예측력이 어떻게 변화하는지 살펴보았다.

TPM = P(에 속하는 관찰값을 잘못 분류) + P(π_2에 속하는 관찰값을 잘못 분류)

$$= p_1\int_{R_2} f_1(x)dx + p_2\int_{R_1} f_2(x)dx \quad (\text{단, } p_i\text{는 } \pi_i\text{에 대한 사전확률}) \quad (1.6)$$

검증집단을 대상으로 모든 구매이력정보를 포함한 경우에 판별분석을 수행한 후, 측정집단과 관측집단간의 오분류 및 정분류 상태를 〈표 6〉과 같은 분류 행렬로 표현

〈표 6〉 분류행렬 결과 (완전모형(full model), 검증집단)

		추정집단		합계
		구매	비구매	
관측집단	구매	829 (43.72%)	1067 (56.28%)	1896 (100%)
	비구매	418 (9.21%)	4119 (90.79%)	4537 (100%)
합 계		1247 (19.38%)	5186 (80.62%)	6433 (100%)

할 수 있다.

한편, 검증집단에서 구매집단과 비구매집단의 사전확률이 각각 $p_1 = 0.2947$, $p_2 = 0.7053$이므로 사전확률을 고려한 총 오분류율은 (1.6)식에 의거하여 $0.2947 \times 0.5628 + 0.7053 \times 0.0921 = 0.2308$로 계산될 수 있다. 역시 동일한 방법으로 측정집단과 검증집단을 대상으로 모든 정보집합을 독립변수로 사용하였을 경우의 오분류율은 다음의 〈표 7〉과 같다.

〈표 7〉을 통해 판별분석을 수행한 후의 오분류율의 변화를 살펴보면 전반적으로 구

〈표 7〉 측정집단과 검증집단에서의 오분류율

	측정집단 (calibration set)	검증집단 (validation set)
base	0.5000	0.5000
demo	0.2947	0.3026
PH1	0.2493	0.2944
PH2	0.2475	0.2721
PH3	0.2358	0.2619
PH4	0.2335	0.2586
PH5	0.2304	0.2572
PH6	0.2299	0.2575
PH7	0.2307	0.2564
PH8	0.2294	0.2564
PH9	0.2298	0.2572
PH10	0.2308	0.2578

매이력정보가 추가됨에 따라 줄어드는 추세를 보이면서 모형의 예측력이 점점 증가하고 있음을 알 수 있다. 또한 측정집단과 검증집단에서 모두 오분류율은 0.5 이하의 값을 가지고 있으므로 아무 정보가 없는 상황(base)보다 월등한 예측력을 나타내고 있음을 알 수 있다.

2. 구매 시점 예측 모형-위험모형

소비재와 같이 반복구매가 일어나는 제품의 경우, 고객이 언제 제품을 다시 구매할 것인지를 예측하는 것은 중요한 이슈이다. 이 때 관심이 되는 변수는 언제 구매가 다시 일어날 것인가 하는 것이다. 즉 다음 구매가 일어나기까지의 지속기간(duration)이 주요 관심의 대상이 되다고 할 수 있다. 지속기간은 변수의 특성상 관측기간 중에서 특정 상태가 발생하지 않는 우측절단(right censored)과 같은 현상이 존재하는 문제를 안고 있기 때문에 일반적인 회귀분석의 기본 가정에는 잘 들어맞지 않는다. 위험모형(hazard model)은 이러한 변수들을 계량적으로 분석하기 위하여 개발된 모형으로 최소제곱법에 의한 추정방법보다 우월한 특성을 갖추고 있어 지속기간의 성격을 지닌 마케팅 변수에 대한 연구들에서 널리 이용되어 왔다. 이런 의미에서 위험모형을 지속기간분석(duration analysis)라고도 일컫는다.

여기서는 구매 시점을 예측하기 위한 구매 간격을 분석하는 모델로서 위험모형을 도입하여 구매 간격을 종속변수로 하는 함수로 모형을 세우고, 구매가 발생하는 위험(hazard)을 예측하는 데 있어 고객의 인구통계정보와 구매이력정보의 조합이 전체 모형의 예측력에 어떻게 영향을 주는 지에 대해 살펴보았다.

확률변수 T_i를 구매 간격 즉, i번째 고객이 가장 최근 구매한 시점인 t_0와 직전 구매

〈표 8〉 위험모형 요약

변 수		위험모형
종속변수(T)	duration	고객의 구매간격
독립변수(X)	gender/income/edu resdol 1-resdol 10 dur 1-dur 10	인구통계변수 - 성별, 소득, 학력 t_0-1기에서 t_0-10기의 구매금액 t_0-1기에서 t_0-10기의 프로모션 반응간격

시점인 t_0-1 사이의 간격 — 이라고 하면, i번째 고객이 어떤 t시점 이전에 구매할 확률을 나타내는 확률분포는 식 (2.1)과 같이 표현될 수 있다.

$$F_i(t) = \int_0^t f_i(s)ds = \Pr(T_i \le t) \tag{2.1}$$

여기서 $f_i(t)$는 확률밀도함수(probability density function)이고 (2.1) 식을 이용하면 i번째 고객이 t시점까지 비구매 상태일 생존함수(survival function)는 $S_i(t) = \Pr(T_i > t) = 1 - F_i(t)$로 표현할 수 있다. 이 경우 위험함수(hazard function)는 i번째 고객에 대한 순간 구매확률(instantaneous probability of exit at time t)를 나타낸다.

$$h_t(t) = \lim_{\Delta t>0} \frac{P(t \le T_i \le t+\Delta t \mid T_i \ge t)}{\Delta t} = \frac{f_i(t)}{S_i(t)} \tag{2.2}$$

즉 위험률 $h_t(t) = -\dfrac{d\ln S_i(t)}{dt}$ 이므로 다음의 관계가 성립한다.

$$f_i(t) = h_i(t)S_i(t) = h_i(t)\exp[-\int_0^t h_i(s)ds] \tag{2.3}$$

그러나 위의 위험모형은 독립변수의 효과를 고려하고 있지 않다는 단점을 안고 있다. 위험모형에서 독립변수 효과를 고려하는 방법으로는 위험함수의 모수를 독립변수의 함수로 가정하여 모형화 하는 방법과 Cox에 의해 개발된 비례위험모형(proportional hazard model)방법의 두 가지가 대표적이다. 본 연구에서는 위험함수(hazard function)를 모형화 하기 위해 전자의 방법인 모수적 회귀분석 방법을 사용하기로 하였다. 한편 Gupta et al.(1994)에서 로그 로지스틱 위험함수(log-logistic harzard)가 구매 간격 분석에 있어 모형의 적합성이나 예측 능력에 있어서 좋은 성과를 낸다는 것이 증명된 바 있으므로 구매 간격에 대한 분포로 로그 로지스틱 분포를 가정하였다. 그러면 어떤 소비자 i가 t시점 이후에 구매를 하게 될 위험함수 $h_i(t)$는 다음과 같은 속도로 감소하게 된다.

$$h_t(t) = \frac{\lambda\gamma(\lambda t)^{\gamma-1}}{1+(\lambda t)^{\gamma}} \tag{2.4}$$

〈표 9〉 위험모형의 추정 계수

변수명		계수추정값	표준오차	Chi-Square	P-Value	
Intercept		3.6202	2.15E-01	14.1500	<.0001	***
구매이력정보	dur1	0.0046	5.98E-04	59.2565	<.0001	***
	dur2	0.0007	6.47E-04	1.1056	0.293	
	dur3	0.0027	5.89E-04	21.0104	<.0001	***
	dur4	0.0006	4.39E-04	1.8046	0.1792	
	dur5	0.0004	4.23E-04	0.9437	0.3313	
	dur6	0.0015	4.20E-04	12.7008	0.0004	***
	dur7	0.0006	3.21E-04	3.4017	0.0651	*
	dur8	0.0006	3.96E-04	2.5299	0.1117	
	dur9	0.0002	2.67E-04	0.4983	0.4802	
	dur10	0.0015	3.74E-04	15.5705	<.0001	***
	resdol1	0.0001	7.19E-03	0.0003	0.987	
	resdol2	-0.0155	1.25E-02	1.5526	0.2128	
	resdol3	-0.0309	9.12E-03	11.5011	0.0007	***
	resdol4	-0.0297	6.15E-03	23.2835	<.0001	***
	resdol5	-0.0196	5.94E-03	10.8272	0.001	***
	resdol6	-0.0083	6.23E-03	1.7930	0.1806	
	resdol7	-0.0189	3.65E-03	26.9268	<.0001	***
	resdol8	-0.0072	5.91E-03	1.4843	0.2231	
	resdol9	-0.0197	5.51E-03	12.7847	0.0003	***
	resdol10	0.0200	4.57E-03	19.0147	<.0001	***
인구통계정보	gender	-0.0230	3.06E-02	0.5635	0.4529	
	income	0.0001	1.51E-03	0.0081	0.9281	
	edu	0.0014	1.99E-03	0.5290	0.467	
scale		0.3818	8.34E-03			

단, ***는 유의수준 0.01 에서 유의, **는 0.05에서 유의, *는 0.1에서 유의

$$\text{단, } \gamma = 1/\sigma,\ \lambda = \exp\{-[\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k]\}$$

측정집단(calibration set)을 대상으로 (2.4)의 모형을 적용하여, 독립변수로 사용되는 정보집합을 변화시키면서 회귀계수벡터 β값을 산출하였다. 이러한 과정을 통해 $t_0 - 10$ 시점의 구매이력정보까지 포함한 완전모형(full model)에서의 회귀계수는 〈표 9〉

와 같이 도출되었다.

한편, 위험함수의 분포가 로지스틱 분포를 따른다고 가정하였으므로 생존함수(survivor function)의 식은 다음과 같다.

$$S(t) = \frac{1}{1+(\lambda t)^{\gamma}} \tag{2.5}$$

(2.5)식을 다음과 같이 변형하면 회귀계수들은 t시점에서 생존할 로그 오즈(log-odds)에 대한 영향력이라고 할 수 있다.

$$\log\left[\frac{S(t)}{1-S(t)}\right] = \beta_0^* + \beta_1^* x_1 + \ldots + \beta_k^* x_k - \gamma \log t \qquad \text{단, } \beta_i^* = \frac{\beta_i}{\sigma} \tag{2.6}$$

(2.6)의 식을 이용하여 결과값들을 해석하면, dur1의 계수는 0.0046이므로 $t_0 - 1$ 시점에서 프로모션에 반응한 기간이 한 단위 증가할수록 t_0 시점에서 구매할 확률에 대한 구매하지 않을 확률의 비는 exp(0.0046/0.3818) = 1.01로서 프로모션에 대해 하루 늦게 반응할수록 해당 시점에서 구매하지 않을 확률이 구매할 확률에 비해 약 1.01배 정도 높아진다는 것을 의미하고 있다.

이와 같은 방법으로 유의한 변수들을 살펴보면, 과거 구매 금액(resdol3, resdol4, resdol5, resdol7, resdol9, resdol10)은 음(-)의 부호를 나타내고, 프로모션에 대한 반응기간(dur1, dur3, dur6, dur7, dur10)은 양(+)의 부호를 나타내므로 과거 구매 금액이 적고 프로모션에 느리게 반응한 고객일수록 구매 간격이 길어진다는 결론을 얻을 수 있다.

추가적으로 σ(scale)의 값은 약0.3818로 추정되었으며 이것은 로그 로지스틱 위험함수의 모양이 뒤집어진 U자 곡선의 모양을 하고 있음을 나타내고 있다. 즉 구매가 일어나게 될 위험은 시간이 지남에 따라 증가했다가 감소하는 추세를 가지게 됨을 알 수 있다.

한편, 이와 같이 측정집단에서 도출된 모형을 검증집단에 적용하여 구매 간격을 예측하고 예측된 값과 실제 구매 간격 사이의 차이를 다음과 같은 식으로 표시되는 RMSE(rooted mean squared error)로 측정하였다.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{T}_i - T_i)^2}{n-1}}$$ $\hat{T}_i$: 예측 구매간격, T_i: 실제 구매간격

구매 간격을 예측하기 위하여 Helsen and Schmittlein(1993)이 제시한 대로 어떤 고객이 t시점에서 생존할 확률(구매가 일어나지 않을 확률)인 $S(t)$의 예측값이 0.5보다 작아지는 시점를 구매가 발생하는 시점으로 보고 그 때의 t 값을 구매 간격의 예측값으로 사용하였다.

〈표 10〉을 통해 위험모형을 수행한 후의 RMSE값의 변화를 살펴보면 전반적으로 구매이력정보가 추가됨에 따라 줄어드는 추세를 보이면서 모형의 예측력이 점점 증가하고 있음을 알 수 있다. 또한 측정집단과 검증집단에서 모두 RMSE값은 아무 정보가 없는 상황(base)에서 평균 반응 간격과의 차이로 계산된 RMSE값보다 월등한 예측력을 나타내고 있음을 알 수 있다.

3. 구매 금액 예측 모형-다중회귀분석

선형회귀분석은 하나의 종속변수와 하나 이상의 독립변수의 관계를 선형으로 가정

〈표 10〉 측정집단과 검증집단에서의 RMSE 값

	측정집단(calibration set)	검증집단(validation set)
base	141.4577	169.2068
demo	125.6388	145.1546
PH1	122.5195	142.3489
PH2	121.7771	121.5026
PH3	118.7373	122.2528
PH4	118.2674	119.5396
PH5	117.3629	121.4494
PH6	117.5312	121.2251
PH7	116.6374	120.0992
PH8	116.4831	121.0167
PH9	116.8779	120.5253
PH10	116.4453	119.7869

〈표 11〉 다중회귀분석 모형요약

변 수		다중회귀모형
종속변수(Y)	targdol	t_0 기의 구매금액
독립변수(X)	gender/income/edu	인구통계변수 - 성별, 소득, 학력
	resdol 1-resdol 10	t_0-1기에서 t_0-10기의 구매금액
	dur 1-dur 10	t_0-1기에서 t_0-10기의 프로모션 반응간격

〈표 12〉 다중회귀분석의 추정 계수

변수명		계수추정값	표준오차	T 값	P-Value	
Intercept		3.0435	2.15E-01	14.15	<.0001	***
구매이력정보	dur1	-0.0045	6.61E-04	-6.83	<.0001	***
	resdol1	0.0492	8.77E-03	5.61	<.0001	***
	dur2	-0.0056	9.15E-04	-6.08	<.0001	***
	resdol2	0.0606	1.50E-02	4.04	<.0001	***
	dur3	-0.0014	6.27E-04	-2.15	0.0315	**
	resdol3	0.0844	1.23E-02	6.88	<.0001	***
	dur4	-0.0035	6.09E-04	-5.79	<.0001	***
	resdol4	0.0149	8.03E-03	1.85	0.0637	*
	dur5	-0.0014	5.74E-04	-2.45	0.0142	**
	resdol5	-0.0019	6.96E-03	-0.27	0.7840	
	dur6	-0.0010	5.46E-04	-1.88	0.0600	*
	resdol6	-0.0012	8.05E-03	-0.15	0.8831	
	dur7	-0.0015	5.15E-04	-2.81	0.0050	***
	resdol7	-0.0021	6.99E-03	-0.30	0.7614	
	dur8	-0.0008	6.81E-04	-1.14	0.2542	
	resdol8	-0.0163	7.93E-03	-2.06	0.0397	**
	dur9	-0.0006	4.32E-04	-1.46	0.1448	
	resdol9	-0.0127	7.03E-03	-1.80	0.0715	*
	dur10	-0.0003	3.89E-04	-0.75	0.4554	
	resdol10	-0.0134	5.21E-03	-2.57	0.0101	**
인구통계정보	gender	-0.0001	4.48E-02	0.00	0.9976	
	income	-0.0006	1.81E-03	0.35	0.7269	
	edu	-0.0010	2.85E-03	-0.33	0.7387	

단, ***는 유의수준 0.01 에서 유의, **는 0.05에서 유의, *는 0.1에서 유의

하여 분석하는 방법으로 마케팅 데이터를 분석하는데 가장 널리 사용되어왔다. 본 연구에서는 연속형 변수(continuous variable)인 구매 금액을 예측하는 모형으로서 선형 회귀분석을 도입하고, 소비자의 최종 구매액을 종속변수로 하였을 때, 이에 영향을 미칠 것으로 예상되는 독립변수들의 정보 집합을 변화시키면서 예측력의 변화를 살펴보았다.

측정집단을 대상으로 다중회귀분석을 사용하여 독립변수로 사용되는 정보집합을 변화시키면서 회귀계수벡터 β값을 산출하였다. 이러한 과정을 통해 t_0-10 시점의 구매이력정보까지 포함한 모형의 회귀계수는 다음과 같이 도출되었다.

〈표 12〉의 결과를 살펴보면 최근의 구매이력정보들이 구매금액 예측에 유의한 영향을 주고 있었으며 인구통계정보는 모두 유의하지 않았다. 또한 유의한 계수들을 중심으로 부호를 살펴보면 과거 구매금액이 많을 수록, 프로모션 받은 이후 구매까지 걸리는 기간이 짧을 수록 미래 구매금액에 양(+)의 영향을 주고 있음을 확인할 수 있었다.

한편, 이와 같이 도출된 회귀식을 검증표본에 적용하여 모형의 예측력의 변화를 살펴보기 위해 본 논문에서는 실제값과과 예측값의 차이를 다음과 같은 식으로 표시되는 RMSE로 측정하였다.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{n-1}} \quad \hat{Y}_i\text{: 예측 구매금액, } Y_i\text{: 실제 구매금액}$$

〈표 13〉을 통해 다중회귀분석을 수행한 후의 RMSE값의 변화를 살펴보면 전반적으로 구매이력정보가 추가됨에 따라 줄어드는 추세를 보이면서 모형의 예측력이 점점 증가하고 있음을 알 수 있다. 또한 측정집단과 검증집단에서 모두 RMSE값은 아무 정보가 없는 상황(base)에서 모집단 전체의 평균 구매 금액와의 차이로 계산된 RMSE값 보다 월등한 예측력을 나타내고 있음을 알 수 있다.

〈표 13〉 측정집단과 검증집단에서의 RMSE 값

	측정집단(calibration set)	검증집단(validation set)
base	1.8490	2.0269
demo	1.7741	1.8981
PH1	1.6399	1.8469
PH2	1.5991	1.8061
PH3	1.5660	1.7868
PH4	1.5618	1.7777
PH5	1.5575	1.7767
PH6	1.5507	1.7761
PH7	1.5523	1.7749
PH8	1.5531	1.7745
PH9	1.5523	1.7743
PH10	1.5526	1.7737

4. 연구의 종합 분석

고객 정보 집합이 달라짐에 따라 각각의 모형의 예측력이 어떻게 달라지는지에 대해 검증집단을 대상으로 분석한 결과를 정리하면 〈표 14〉와 같다.

〈표 14〉에서 알 수 있듯이 고객에 관해 아무런 정보를 가지고 있지 않고 무작위로 마케팅을 수행했을 때 보다 기본 인적 정보를 이용하여 타겟 마케팅을 할 경우에 타겟팅의 정확도가 증가하는 것을 확인할 수 있다. 기본 인적 정보 외에 구매 이력 정보를 1개라도 포함한 모형을 사용했을 경우, 고객의 미래 구매 여부, 구매 시점, 구매 금액에 관한 예측력이 상승했음을 확인할 수 있었다. 이로써 구매 정보가 인적 정보에 비해 구매 행동을 예측하는데 보다 유용할 것이라는 결론을 내릴 수 있다.

한편, 구매 이력 정보 추가에 따른 예측력의 변화와 개선 정도를 그래프로 나타내면 〔그림 3〕~〔그림 5〕과 같다.

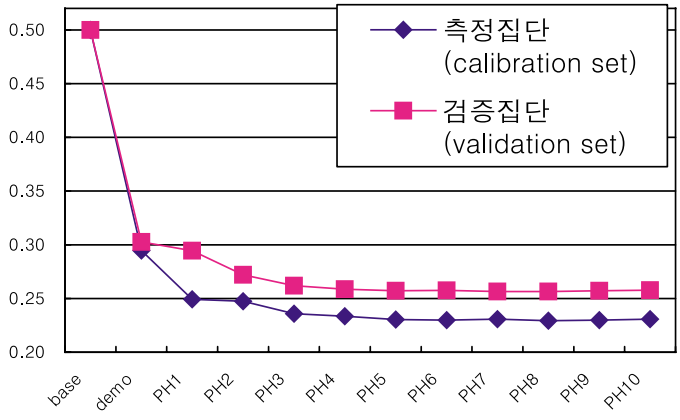
세가지 모형에서 모두 구매이력정보가 추가될수록 예측력이 증가(또는 오분류율/RMSE 값이 감소)하고 있음을 확인할 수 있었다.

한편 구매이력 데이터가 증가하면서 측정집단의 예측력은 점차 증가하는 추세를

〈표 14〉 예측 모형과 정보 집합의 변화에 따른 예측력의 변화

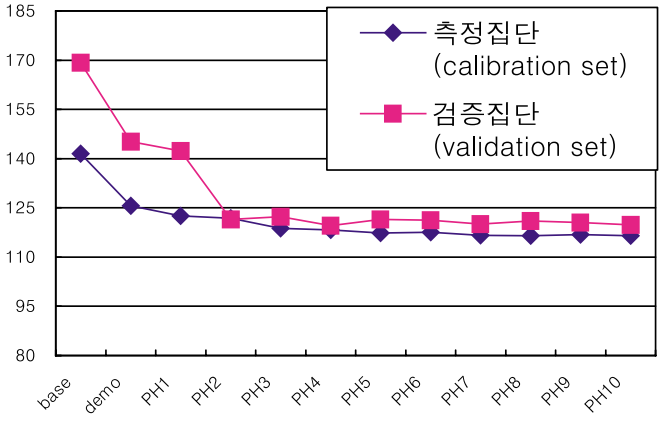
	구매여부(오분류율)	구매시점(RMSE)	구매금액(RMSE)
Base	0.5000	169.2068	2.0269
Demo	0.3026	145.1546	1.8981
PH1	0.2944	142.3489	1.8469
PH2	0.2721	121.5026	1.8061
PH3	0.2619	122.2528	1.7868
PH4	0.2586	119.5396	1.7777
PH5	0.2572	121.4494	1.7767
PH6	0.2575	121.2251	1.7761
PH7	0.2564	120.0992	1.7749
PH8	0.2564	121.0167	1.7745
PH9	0.2572	120.5253	1.7743
PH10	0.2578	119.7869	1.7737

(단, PH1-PH10은 구매이력정보가 하나씩 추가된 정보집합)

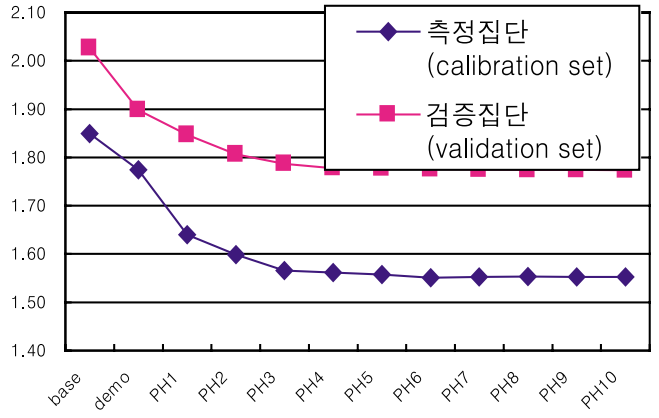


〔그림 3〕 구매 여부 예측: 판별분석의 오분류율

보이나 검증집단의 경우 불규칙적으로 감소하는 부분이 있는데, 이것은 측정집단에서 모형이 과추정(over-fitting)된 결과라 볼 수 있다. 이를 통하여 구매이력정보가 한 단위 추가될 때 예측력은 증가하되 한계 예측력은 점차 감소한다는 결론을 내릴 수 있다.



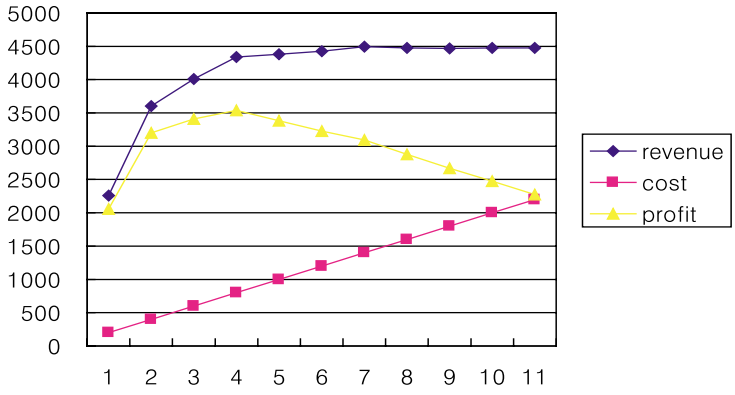
〔그림 4〕 구매 시점 예측: 위험모형의 RMSE



〔그림 5〕 구매 금액 예측: 다중회귀분석의 RMSE

3. 구매 이력 데이터의 자산으로서의 가치평가

아래에서는 실증분석에서 도출된 결과를 바탕으로 고객 구매 이력 데이터 획득 비용을 고려한 일대일 마케팅의 수익성 분석 방법을 제시하고자 하였다. 검증집단을 이용한 구매금액 예측모형의 측정에서, 예측력이 한 단위 증가할 때 구매금액도 비례



〔그림 6〕 타겟팅 효과: 비용함수가 증가하는 경우

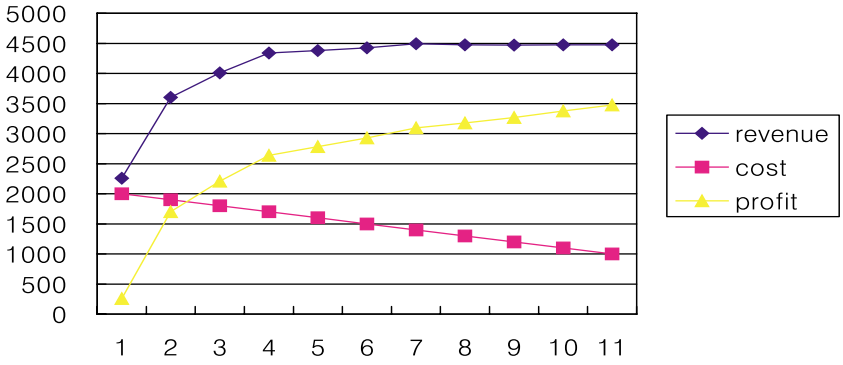
하여 증가한다는 결과를 전제로 하였다. 한편, 고객 정보를 한 단위 얻는데 드는 비용은, 선형적으로 증가하는 경우와 감소하는 두 가지 경우를 상정하여 수익(revenue), 비용(cost), 이익(profit)을 나타내는 그래프를 도시하였다.

- 수익: 타겟팅의 효과로 얻게 되는 기업의 기대 수익
- 비용: 고객 정보 획득 비용+유지 비용
- 이익: 타겟팅의 효과로 얻게되는 기업의 기대 이익 (수익 – 비용)

1) 고객 정보 획득 비용이 증가할 경우

고객과의 거래 정보를 얻는 비용이 모든 거래에 대해 동일한 산업군의 경우, 고객의 수가 증가할수록 비용은 그에 비례하여 선형적으로 증가하게 된다. 이런 경우 타겟팅 효과로 얻게 되는 수익을 그래프로 나타내면 〔그림 6〕과 같은 결과를 얻게 된다.

즉, 타겟팅의 효과로 얻게 되는 기업의 기대 수익은 모델의 예측력이 높아질수록 증가하게 될 것이다. 이에 고객의 구매이력정보를 획득하는 데 드는 비용은 데이터 개수에 선형 비례한다고 가정할 때, 한계 수익이 한계 비용과 같아지는 지점(MR = MC)이 존재함을 알 수 있다. 즉, 기업의 입장에서 고객 데이터를 획득하는 비용을 고려할 때, 타겟팅의 효과로 얻게 되는 수익이 가장 높아지는 최적 지점이 존재한다



〔그림 7〕 타겟팅 효과: 비용함수가 감소하는 경우

고 말할 수 있다.

2) 고객 정보 획득 비용이 감소하는 경우

신용카드 산업과 같이 경우, 초기에는 고객 정보를 획득하는 데 많은 비용이 소모되는 반면 어느 정도 기반이 구축이 되면 고객의 거래 정보가 회사의 내부에 자연스럽게 축적되면서 고객 정보를 얻는 추가 비용이 현저하게 줄어드는 추세를 가지게 된다. 이런 경우를 상정하여 위와 같은 방법으로 이익을 산출해 보면 〔그림 7〕과 같이 데이터의 개수에 비례하여 증가하는 형태를 보이므로 거래 데이터를 많이 활용하면 할수록 높은 이익을 얻을 수 있을 것이다.

V. 요약 및 결론

본 연구에서는 CRM의 확산으로 그 중요성이 더해지고 있는 고객의 구매 이력 데이터의 가치를 계량적으로 측정하여 고객 데이터를 자산으로서 평가하는 것에 목적을 두었다. 이를 위해 고객의 구매 여부 예측 모형, 구매 시점 예측 모형, 구매 금액 예측 모형 등 세 가지의 모형에서 여러 가지 조합의 정보집합을 설명변수로 사용하여 데이터 종류와 개수의 변화에 따른 예측력의 개선 정도를 살펴보고자 하였다.

모형에 적용한 결과 예측에 사용된 구매 이력 데이터의 개수가 증가할수록 모형의

예측력이 공통적으로 증가하는 것을 알 수 있었다. 단 증가의 정도는 체감 하는 것으로 나타나 구매 이력 데이터의 단위 가치가 점차 감소한다는 결과를 확인하였다. 실무에 적용할 경우 주어진 고객 정보 획득 비용에 대하여 고객 정보의 가치가 극대화되는 지점이 있음을 유추할 수 있다. 현업에서 이러한 타케팅 전략을 수행하려고 할 때 고객 데이터의 자산 가치 평가에 이러한 점을 반영해야 할 것이다.

본 연구는 고객 구매 이력 데이터의 가치를 평가하는 기준으로 모형의 예측력을 선택하였다. 예측력이 높을수록 모형의 적합도가 높아져 구매 확률이 높은 고객을 정확히 선별하여 일대일 마케팅을 수행할 수 있기 때문에 기업의 현금흐름을 상승시킬 것이라는 가정을 하였기 때문이다. 그러나 산업의 특성이나 고객들의 이질성(heterogeneity)이 반영되지 않았고, 예측력이 모형을 평가하는 유일하고 절대적 기준이 될 수 없다는 한계를 가진다. 또 데이터가 추가됨에 따른 전체적인 모형의 개선 정도에 초점을 맞추어 독립 변수들간의 자기상관(auto-correlation)현상을 통제하지 못했다는 한계가 있다.

이에 더하여 다양한 구매 예측 모형들 모두가 아닌 특정 모형들에 연구를 국한시켰으므로 결과를 일반화하기에는 무리가 있을 수 있으며, 모형의 과대 적합으로 인한 예측력의 저하라는 요인을 통제하지 못했던 한계가 있다. 향후에는 산업의 특성을 반영하여 데이터의 가치를 측정하는 새로운 기준을 정하고 실제 현업 데이터를 바탕으로 고객 데이터의 가치를 보다 정교할 모형으로서 평가하는 연구가 필요할 것이다.

참고문헌

김병도 (2002). *CRM 전문가를 위한 통계 방법론*. 서울대학교 경영대학.

박찬욱 (1996). *데이터베이스 마케팅*. 연암사.

성웅현 (1997). *응용다변량분석*. 탐진.

송문섭, 조신섭 (2002). *SAS를 이용한 통계자료분석*. 자유아카데미.

이유재, 최정환 (2001). "죽은 CRM 살아있는 CRM". *한언*, 2001. 8.

한상만, 박승배, 정남호 외 (2000). "인공신경망과 로짓모형을 이용한 내구재의 구매의도 예측에 관한 비교연구." *마케팅연구*. 2000. 9, pp. 71-92.

Banslaben, J. (1992). "Predictive Modeling." in E. Nash (eds.), The Direct Marketing Handbook, McGraw-Hill: New York, NY.

Bass, F. M. (1969). "A New Product Growth Model for Consumer Durables." *Management Science*, 15, 215-227.

Bawa, Kapil and Robert W. Shoemaker (1987). "The Effects of a Direct Mail Coupon on Brand Choice Behavior." *Journal of Marketing Research*. 24, 370-376.

Bitran, G.R. and S.V. Mondschein (1995). "An application of yield management to the hotel industry considering multiple day stays." *Oper Res* 43(3), 427-443.

Blattberg, Robert C., Gary Getz, Jacquenlyn S. Thomas (2001). "Customer Equity: Building and Managing Relationships as Valuable Assets." Harvard Business School Press.

Blattberg, Robert C. and John Deighton (1991). "Interactive Marketing: Exploiting the Age of Addressability." *Sloan Management Review*, 33(1): 5-14.

Bult, J. (1993). "Semiparametric Versus Parametric Classification Models: An Application to Direct Marketing." Journal of Marketing Research, 30, 380-390.

Bult, J. and Wansbeek, T. (1995). "Optimal Selection for Direct Mail." *Marketing Science*, 14, 4, 378-394.

Cox, D. (1972). "Regression Models and Life Tables (with discussion)." Journal of the Royal Statistical Social Bulletin, 34, 187-220.

Colombo, R. and Jiang, W. (1999). "A Stochastic RFM Model." *Journal of Interactive Marketing*, 13, 3, 2-12.

DeSarbo, W.S., & Ramaswamy, V. (1994). "CRISP: Customer Response Based Iterative Segmentation Procedures for Response Modeling in Direct Marketing." *Journal of Direct Marketing*, 8 (3), 7-20.

Dunn, R., S. Reader and N. Wrigley (1983). "An Investigation of the Assumptions of the NBD Model as Applied to Purchasing at Individual Stores." *Applied Statistics*, 32, 249-259.

Girard, P. (2000). "The Business of Selects." *Catalog Age*, May, 33.

Greene, J.D. (1982). Consumer Behavior Models for Non-Statisticians, New York: Praeger.

Gupta, Sachin and Pradeep K. Chintagunta (1994). "On Using Demographic Variables to

Determine Segment Membership in Logit Mixture Models." *Journal of Marketing Research*, 31 (February).

Helsen, Kristiaan and David C. Schmittlein (1993). "Analyzing Duration Times in Marketing: Evidence for the Effectiveness of Hazard Rate Models." *Marketing Science*, Vol. 11, No. 4. Fall 1993.

Kiefer, N. (1998). "Economic Duration Data and Hazard Functions." *Journal of Economic Literature*, 26, 646-679.

McEachern, C. (1998). "Redefining Your Best Customers." *Catalog Age*, May, 111.

Miller, P. (1998). "RFM, to RMP, to RMPF." *Catalog Age*, April 1, 47.

Paul D. Allison (1995). "Survival Analysis Using The SAS System: A Practical Guide." SAS Institute Inc.

Rossi, Peter E., Robert E.McCulloch and Greg M. Allenby (1996). "The Value of Purchase History Data in Target Marketing." Marketing Science, 15, No.4, 321-340.

Schmid, J. and Weber, A. (1995). "Catalog Database Marketing Done Right." *Target Marketing*, 18, 10, 34-37.

Schimittlein, D. and Peterson, R. (1994). "Customer Base Analysis: An Industrial Purchase Process Application." *Marketing Science*, 13, 1, 41-65.

Schimittlein, D.C. and V. Mahajan (1982). "Maximum Likelihood Estimation for an Innovation Diffusion Model of New Product Acceptance." *Marketing Science*, 1, 57-78.

Schimittlein, D.C. and D.G. Morrison (1983a). "Modeling and Estimation Using Job Duration Data." *Organizational Behavior and Human Performance*, 32, 1-22.

Schimittlein, D.C. and D.G. Morrison (1983b). "Prediction of Future Random Events with the Condensed Negative Binomial Distribution." *Journal of the American Statistical Association*, 78, 449-456.

Stone, B. (1996). Successful Direct Marketing Methods. Chicago: NTC Business Books.

Wedel, Michel, Wayne S. DeSarbo, Jan Roelf Bult, and Venkatram.

How do types and number of the customer data affect the goodness-of-fit of the customer purchase models?

Yun Kyung Oh*
Jikyung Kim**
Sang-Hoon Kim***

ABSTRACT

The advancement of computer and telecommunicaton technology is providing numerous ways for the marketing managers to collect and utilize consumer contact data. In other words, marketers can now escape the limited boundaries of mass marketing channels and begin to address individual needs of each customer, thereby realizing one-to-one marketing. Those in academic fields have worked hard to develop models that predict the purchase pattern and identify the causal factors. These efforts can be put into one word that has become quite a sensation: the CRM(customer relationship marketing).

In this study, We have tried to quantitatively measure the value of customer information, which has increasingly become important with the growing use of one-to-one marketing. The customer information has been categorized into two, namely demographic information and purchase history information. Different number and type covariates have been incorporated into several models to see how they contribute to improving the goodness-of-fit

*Ph. D candidate Purdue University : oh13@purdue.edu
**Master of Business Administration, Seoul National University : jeanne@hanafos.com
***Assistant Professor of Marketing, Seoul National University : profkim@snu.ac.kr

of the models.

The results of the study are as follows.

First, even limited amount of demographic information about the customer can drastically improve the goodness of fit of the models for target marketing.

Second, consumer purchase history data is more effective and better serve the purpose than the demographic data when predicting purchase behavior.

Third, as more purchase history data are added, more accurate the model becomes.

Lastly, the marginal improvement of predictive capability of the model tends to decline as more information is added. Therefore it can be deduced that there exists an ideal point in the projectile of marketing budget that compromises cost and benefit.

Key Words: CRM, direct marketing, customer data