코호넨네트워크와 생존분석을 활용한 신용 예측

하성호* · 양정원**† · 민지홍**

Credit Prediction Based on Kohonen Network and Survival Analysis

Sung Ho Ha* · Jeongwon Yang** · Jihong Min**

■ Abstract ■

The recent economic crisis not only reduces the profit of department stores but also incurs the significance losses caused by the increasing late-payment rate of credit cards. Under this pressure, the scope of credit prediction needs to be broadened from the simple prediction of whether this customer has a good credit or not to the accurate prediction of how much profit can be gained from this customer. This study classifies the delinquent customers of credit card in a Korean department store into homogeneous clusters. Using this information, this study analyzes the repayment patterns for each cluster and develops the credit prediction system to manage the delinquent customers. The model presented by this study uses Kohonen network, which is one of artificial neural networks of data mining technique, to cluster the credit delinquent customers into clusters. Cox proportional hazard model is also used, which is one of survival analysis used in medical statistics, to analyze the repayment patterns of the delinquent customers in each cluster. The presented model estimates the repayment period of delinquent customers for each cluster and introduces the influencing variables on the repayment pattern prediction. Although there are some differences among clusters, the variables about the purchasing frequency in a month and the average number of installment repayment are the most predictive variables for the repayment pattern. The accuracy of the presented system reaches 97.5%.

Keyword : Kohonen Network, Clustering, Cox Proportional Hazard Model, Credit Prediction System

논문접수일 : 2008년 10월 07일 논문게재확정일 : 2009년 03월 27일
논문수정일(1차 : 2008년 12월 28일, 2차 : 2009년 03월 23일)
* 경북대학교 경상대학 경영학과
** 경북대학교 일반대학원 경영학과
† 교신저자

1. 서 론

한국 유통시장의 중심에 있던 백화점은 1990년대 이후 할인점과 홈쇼핑, 온라인 쇼핑몰의 활성화로 유통시장에서의 비중이 감소되고 있고 2000년대 들어 계속된 소비위축으로 인한 경기악화로 백화점 카드의 부실화 위험이 높아지면서 이에 대한 대비책이 요구되고 있다. 이와 관련하여 신용예측에 관련된 기존의 연구를 살펴보면, 우량, 잠재적 불량, 불량으로 삼분한 모형도 있기는 하지만[4], 대부분의 경우 불량채권 발생에 초점을 맞추어 우량·불량고객으로만 군집을 나누어 예측하는 이분법적인 모형을 사용하였고, 회복가능고객군집, 즉 연체상태에서 정상고객이 되는 군집에 대해서는 고려하지 않았다[13, 18, 21, 32, 39, 45]. 또한 신용예측의 목적이 단순한 부실여부의 추정에서 고객을 통한 수익성 증대로 변화함에 따라 부실로 인한 손실의 추정과 더불어 수익에 미치는 영향들도 추정할 수 있는 모델이 필요하게 되었다[7].

따라서 본 연구에서는 회복가능고객 군집에 초점을 맞추어 연체기록이 있는 고객 중 회복가능고객의 데이터를 분류하여 군집을 만들고 각 군집별로 고객의 연체탈출 유형을 분석하여 백화점 연체고객관리를 위한 시스템을 제안하고자 한다. 본 연구의 연체고객 신용예측 시스템을 통해 제공될 세분화된 고객군집의 기간별 회수가능 불량채권에 대한 정보는 백화점 신용판매 관리부서에서 고객군집별로 차별적인 연체 대응전략을 수립하고, 연체관리에 효율성을 높이는 도움을 줄 수 있을 것으로 기대된다. 본 연구에서는 연체상태에서 정상고객으로 신용회복이 된 고객들을 대상으로 데이터마이닝 기법인 코호넨 네트워크와 의학적인 통계 기법인 콕스의 비례적 위험모형을 결합하여 연체고객의 유형을 분류하고 연체탈출의 예상기간과 연체탈출에 영향을 주는 요인을 함께 고려할 수 있는 신용예측 시스템을 제안하고자 한다.

본 연구의 구성은 다음과 같다. 제 1장에서는 본 연구의 배경과 목적을 언급하고 제 2장에서는 신용예측 시스템의 필요성과 유형에 대해서 알아본 후 다양한 기법으로 구현된 기존의 신용예측 시스템에 대한 연구결과를 살펴본다. 제 3장에서는 본 연구에서 연체고객 신용예측 시스템 구현을 위해 사용하고자 하는 모형의 프레임워크를 제시한다. 제 4장에서는 본 연구의 모형을 실제 기업의 데이터 입력을 통해 분석하고 결과를 도출하여 연체고객 신용예측 시스템 모형을 구축하고 이를 검증하며 로지스틱 회귀 모형과 이를 비교함으로써 모형의 우수성을 검증한다. 또한 검증된 모형을 통해 의미있는 경영정보를 도출한다. 제 5장에서는 결론 부분으로 본 연구의 결과를 요약하고, 그 의의와 한계점에 대해서 살펴본다.

2. 신용예측 시스템의 이론적 배경

2.1 신용예측 시스템의 필요성

금융기관들은 대량의 고객 데이터 분석을 통해 서비스 개선, 기존고객의 이탈방지, 정확한 신용예측을 통한 대출, 신상품의 개발 등의 노력을 기울여 왔다. 특히, 신용예측 시스템은 우량고객과 불량고객의 판별 및 고객 신용등급의 관리에 활용되어 불량채권 발생률을 미연에 감소시키는 역할을 담당했다. 또한 신용예측을 위해 수집된 자료들과 그 결과를 바탕으로 고객에 따라 차별화 된 금융상품과 여러 가지 혜택을 제공하고 위험상황을 사전에 통지하는 등의 고객 관계마케팅을 실현하는데 도움을 주었다. 이와 같은 이유로 신용예측 시스템은 기업의 수익을 증대시켜주기 때문에 금융회사의 운영에 있어서 필수적인 부분이라 하겠다[6].

따라서 금융기관들은 대출자산이 위험에 노출되는 정도를 줄이고 연체 가능성이 있는 잠재적 위험고객을 선별하여 관리할 수 있는 전략과 시스템 구축으로 대출 손실을 최소화 할 필요가 있다. 이러한 사전적 신용예측 시스템은 금융회사의 여신기능에 다음과 같은 긍정적인 효과를 가져다준다. 첫째, 금융회사의 대출심사에 객관적인 정보를 제공함으

로써 최적의 대출을 결정할 수 있게 한다. 둘째, 우량·불량 고객을 선별함으로써 고객의 신용도에 따라 신용사용을 조정함으로써 연체율과 부실채권을 감소시킬 수 있다. 셋째, 대출고객들의 신용정도에 의해 신용한도를 효과적으로 분배함으로써 금융자금을 보다 효율적으로 배분할 수 있다. 넷째, 시스템 초기 개발비용이 많이 소요되지만 개발 후 신용조사비용이 줄어들어 전체적인 여신관리 비용을 절감할 수 있다.

2.2 신용예측 시스템의 유형

소비자금융에서 신용예측 시스템은 신용평점시스템(Credit Scoring System)과 행동평점 시스템(Behavior Scoring System), 상환평점 시스템(Recovery Scoring System) 및 생존분석(Survival Analysis) 방법이 있다[7].

신용평점 시스템은 고객에 대한 통계적 정보를 이용하여 고객의 신용 정도를 평점의 형태로 계량화하여 분석하는 시스템으로[36, 31, 41], 다변량 통계분석(판별분석, 로지스틱 회귀모형 등) 및 인공지능과 같은 모형이 활용되고 있다[21, 44, 33, 49].

행동평점 시스템은 이자 납부 등 고객들의 대출 이후 사후적인 거래정보를 분석하여 고객의 상환연기나 재대출조건을 차등 적용하는 등, 신용관리를 목적으로 하는 시스템이다[13, 28, 46]. 행동평점시스템의 경우 신용상태 변화를 추적하기 위해 거래내역이 포함된다는 점에서 신용평점 시스템과 차이를 보이지만[45], 접근 방식에는 차이가 없음에도 대부분의 학술적인 연구는 신용평점 시스템에 집중되어 왔고, 최근 들어 데이터마이닝 기법이 발달됨에 따라 행동평점 시스템에 대한 학술적 관심이 높아지고 있다.

상환평점 시스템은 행동평점 시스템의 일종으로서 연체가 발생하면 이를 점수화하여 부실채권 회수율을 추정할 수 있는 시스템이다. 생존분석은 신용평점 시스템의 로지스틱 회귀모형을 발전시킨 것으로 특정시점에 대출자산의 위험을 반영하여 수익을 추정할 수 있는 시스템이다. 신용평점 시스템은 부도율의 가능성을 추정할 수 있는 모델이지만 생존분석모델은 특정 시점에서 부도율을 추정하고 이를 이용하여 그 시점에서의 위험이 반영된 수익성의 추정을 가능하게 한다.

2.3 신용예측 시스템 구현 기법

신용예측 시스템의 대표적인 예로 전통적으로 연구에 사용되어온 모형으로는 다변량 판별분석이나 로지스틱 회귀분석, 프로빗 분석과 같은 통계학적 모형[24, 48, 39]와 선형계획법[40]과 같은 경영과학 모형을 들 수 있다. 최근 들어서는 의사결정나무, 인공 신경망 등의 인공지능 모형을 이용한 연구가 활발하게 진행되었는데, 특히 인공신경망 모형을 이용한 연구가 좋은 결과를 보여주고 있다[35, 47].

2.3.1 통계학적 기법

초기 신용위험관리 시스템은 전통적 통계기법을 이용하였는데 다변량 회귀분석은 James[36]에 의해 신용평점 시스템으로 구현되었고, 로지스틱 회귀분석은 선형모형의 단점을 극복하기 위해 목표변수가 이진변수일 때 사용하는 특수형태의 회귀분석 기법으로 개발되어 이를 통해 필요한 예측을 하거나 통계적 추론을 하게 되었다[16]. 판별분석 모형은 관찰된 자료의 어떤 특성을 바탕으로 관찰 값을 두 개 이상의 그룹에 각각 구분되도록 하여 주어진 상황에서 응답자들의 행동을 예측하는 것이다[8].

최근에는 생존분석모형을 사용하여 구축된 신용예측 시스템도 제안되고 있다[10, 27]. 생존분석 모형은 어떤 사건이 발생할 때까지의 시간으로 자료가 주어진 경우 이를 분석하는 통계적 방법으로서 사건의 발생 여부에 대해 불확실한 자료가 포함되어 있다는 특징을 가지고 있으며 신용예측을 위해서는 중도 절단된 자료의 처리가 가능한 로지스틱 회귀분석의 일종인 Cox의 비례적 위험 모형(Cox proportional hazard model)이 많이 사용된다[12, 14, 43].

2.3.2 데이터마이닝 기법

신용예측 시스템을 위해 주로 사용되고 있는 데이터마이닝 기법으로는 인공 신경망, 의사결정나무, 유전자 알고리즘 등이 있다. 데이터마이닝에서 가장 중요한 영역들 중의 하나인 분류[15]를 기본으로 하는 의사결정나무 모형은 과거에 수집된 데이터를 분석, 이들 사이에 존재하는 패턴의 분류모형을 나무의 형태로 만드는 것으로 고객을 성격에 따라 우량, 불량고객으로 쉽게 분류할 수 있기 때문에 신용예측에서 널리 사용되고 있는 기법이다[17, 34]. 또한 많은 요인들을 토대로 의사결정을 내릴 필요가 있을 때, 어떤 요인이 고려 대상이 되는지를 구별하는데도 도움을 준다[42].

인공신경망모형은 통계적 가설이 필요 없으면서도 비선형적인 회귀모형을 설명하기에 적당하기 때문에 신용예측에서 뛰어난 성과를 보여 주고 있다[11, 19, 22, 23]. 유전자 알고리즘 모형은 인공지능의 한 모형으로 자연계에서 생물의 유전과 진화의 메커니즘을 공학적으로 모델화하는 것에 의해 생물이 갖는 환경에서의 적응능력을 응용한 자연도태의 원리에 근거를 둔 최적화 기법이다[30]. 2차원 이상의 복잡한 탐색공간에서 전 범위의 최적해(global optimal solution)를 탐색하는데 매우 효율적이라고 증명되어져 왔다[26].

이상의 데이터마이닝 기법을 이용한 신용예측 시스템에 대한 기존 연구를 통해 어떤 기법이 가장 뛰어난지 절대적인 판단이 어렵다고 여겨지는[45] 가장 근본적인 이유는 과학습(overfitting), 최적모형 설정의 어려움 또는 학습방식의 부적절성 등과 같은 문제 때문에 광역최적(global optimum)을 보장하지 못했기 때문으로 단일 모형에 의존하는 기법이 가지고 있는 한계점이라 할 수 있다[29]. 따라서 연구 주제에 대한 최적기법을 찾기 위해서는 각 단일 모형의 장점들만을 취하여 최적의 통합 신용예측 모형을 구성하는 것이 보다 효율적일 것으로 생각된다[1, 2, 37].

2.3.3 기법의 혼합

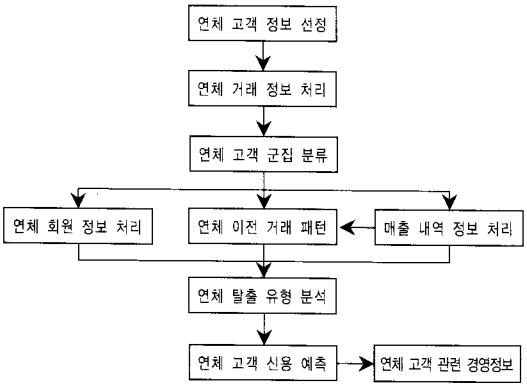
최근의 문헌들을 보면 데이터마이닝 기법과 통계적 기법 모형들을 비교 분석하여 예측률이 높은 모형을 제안하거나 혹은 이들의 혼합모형을 제안하고 있다. 김갑식은 다계층 퍼셉트론과 다변량 판별분석, 그리고 의사결정나무 및 로지스틱 회귀분석을 적용하여 각각의 개별 모형을 도출하고 이를 유전자 알고리즘을 이용하여 통합한 최종 모형을 구해 그 결과를 각 단일 모형과 비교하였다[1]. Chen과 Huang은 결과 해석이 어려운 신경망의 단점을 보안하기 위하여 유전자 알고리즘을 이용하여 신경망 예측결과 '거절'로 나오면 유전자 알고리즘을 이용하여 변수들의 조건을 변화시킨 후 '신용승인'이 될 수 있는 조건을 찾아내 거절사유를 설명하였다[18]. Lee는 판별분석, 로지스틱 분석 그리고 신경망을 비교하여 신경망의 예측률이 가장 좋았다고 했는데 판별분석에서 선택된 변수를 이 신경망의 입력변수로 사용하는 혼합모형을 구현할 경우 신경망보다 예측률이 높았다고 했다[39].

본 연구에서 제안하고자 하는 신용예측 시스템도 예측률을 높이기 위하여 데이터마이닝의 코호넨 네트워크로 먼저 군집을 분류하고 통계학적 기법인 생존분석 중 비례적 위험모형을 이용하여 분석 모형을 생성하는 혼합 모형으로 상환평점시스템의 일종이라 할 수 있을 것이다.

3. 연체고객 신용예측 시스템 개발의 프레임워크

본 연구는 D백화점 신용카드 연체고객들의 연체정보를 이용하여 신용예측 시스템을 개발하는 것을 목표로 한다. 앞서 언급되었듯이 기존의 연구에서는 회복가능고객을 대상으로 한 분석은 찾기 힘들었고, 방법론에서도 경영학적 영역에서 생존분석이나 비례적 위험모형을 이용하여 구현된 모형도 드물다. 따라서 본 연구에서는 연체 고객 군집 분류를 위한 데이터마이닝 군집분류기법인 코호넨 네트워크와 여기에서 분류된 연체 고객 군집별로 연체

탈출유형 분석을 위해 콕스의 비례적 위험모형을 결합한 모형을 제시하고자 한다. 본 연구에서 제안할 연체고객 신용예측시스템의 프레임워크는 연체고객 정보 획득 및 전처리, 연체 고객 군집분류, 연체고객 군집별 연체 탈출 유형 분석, 그리고 연체 고객 신용예측시스템 적용 단계로 나눠 볼 수 있다.



[그림 1] 연체고객 신용예측 시스템 개발의 프레임워크

3.1 연체고객 정보획득

본 연구의 자료는 2003년 D백화점의 신용카드 연체기록이 있는 고객들의 정보를 D백화점에서 받은 것으로 회원정보 50,496건, 연체정보 1,367,506건, 매출내역정보 13,561,909건으로 구성되어 있다. 본 연구에서는 연체상태에 들어간 고객이 그 연체상태에서 탈출하는 유형을 분석하는 것이 목적이며, 미상환 고객은 본 연구에 사용된 데이터에서 차지하는 비율이 0.4% 정도로 작고, 또한 이들은 악성채무자로서 특별히 따로 관리해야할 필요가 있으므로 연체 고객 군집 분류 단계에서부터 상환 정보가 없는 악성 고객은 분석에서 제외한다.

획득된 연체고객 관련정보는 분석 단계 전에 입력변수 생성을 위해 전처리를 하게 된다. 총 연체회원 중 정보가 부정확한 8,501명을 제외한 41,831명의 연체고객 정보가 연체고객 유형 분류를 위해 사용되었고 세부 내용은 아래 <표 1>과 같다. 연체 거래 정보는 연체 거래 테이블에 입력된 41,831

〈표 1〉 연체 회원 정보

칼 럼	내 용	칼 럼	내 용
MEM_NO	회원번호	SIDO	자택주소(시)
AGE	성별	SIGUN	자택주소(구)
SEX	나이	DONG	자택주소(동)
TRADE_SUSP	거래상태(0 : 정상, 1 : 정지)	MEM_REGI_YMD	회원등록일
BIRTH_DAY	생년월일	CARD_ISSUE_YMD	카드발급일
MARRI_YMD	결혼기념일	VALID_YMD	유효기간
POST_NO	자택우편번호	JOB_NM	직업명

〈표 2〉 연체 거래 정보

칼 럼	내 용	칼 럼	내 용
DELAY_MON	연체개월	NORM_DEMAND_AMT	정상청구 원금 금액
SALE_FG	할부구분(0 : 일시불, 1 : 할부)	DELAY_PRIME_AMT	연체청구 원금 금액
ACK_NO	승인번호	DELAY_FEE_AMT	연체청구 수수료 금액
DEMAND_YM	청구연월(대상 연월)	DELAY_ADD_AMT	연체청구 이자 금액

〈표 3〉 매출 내역 정보

칼 럼	내 용	칼 럼	내 용
PURCH_YMD	매출일자	UGOODS_NM	상품명
PURCH_STORE_CD	매출사업장	PURCH_MD	매장구분코드
SEQ_NO	순번	MD_NM	매장명
GOODS_CD	상품코드	PURCH_AMT	매출금액

건의 자료로 <표 2>와 같은데 연체 거래에 대한 분석은 월 단위로 이루어지게 된다. 매출내역정보는 D백화점 신용카드 고객들 중, 임의로 선택된 160,371명에 대한 2003년의 매출내역정보를 추출한 자료로 이중 21,464명이 2003년 연체한 기록이 있었다. 따라서 매출내역정보가 있는 21,464명의 매출내역 정보가 연체 회원 정보와 함께 유형별로 분류된 연체 탈출 분석을 위해 사용 되었다.

3.2 연체고객 군집분류

3.2.1 코호넨 네트워크(Kohonen Network)

본 연구에서 연체고객의 유형을 분류하여 군집화하기 위해 코호넨 네트워크를 사용한다. 코호넨 네트워크는 입력 데이터의 특징을 추출하여 경쟁학습을 통해 지도를 형성함으로써 미지의 데이터 패턴을 인식할 수 있어 분류에 유용한 신경망의 일종이다[3, 38].

입력을 위한 데이터의 선정과 처리가 끝나면 그 자료를 입력변수로 사용하여 군집 분석을 진행하게 된다. 예측력을 높이기 위해 전체 연체 고객을 각각의 특성에 따라 군집을 나누고 그 군집 내에서 연체 상환 능력을 비교하게 된다. 코호넨 네트워크에서 어느 정도 군집이 이루어지고 나면 경쟁층의 노드가 서로 가까울수록 입력벡터들이 서로 유사하다는 특징을 갖게 된다. 단순히 군집화만 하는 것이 아니라, 이러한 topographic order를 유지한다는 점이 코호넨 네트워크가 여타의 다른 군집화 방법에 비해 갖는 장점이라고 할 수 있다. 본 연구에서는 코호넨 네트워크의 이러한 성질을 이용하여 의미 있는 차이를 보이는 군집을 형성하고자 한다. 분석을 위한 툴로 SAS Enterprise Miner 6.0의 SOM/Kohonen 노드를 사용한다[9].

3.3 연체고객 군집별 연체탈출유형 분석

3.3.1 콕스의 비례적 위험모형(Cox Proportional Hazard Model)

본 연구에서 연체 고객의 연체 탈출 유형을 분석하기 위해 생존분석, 그 중에서도 콕스의 비례적 위험 모형(Cox Proportional Hazard Model)을 사용한다. 비례적 위험 모형이란 시간 개념이 들어간 로지스틱 회귀분석으로 생존 시간에 관한 자료를 분석하는 생존분석과 관련된 통계적 기법으로 여러 예후 변수들의 영향을 동시에 알아보는 다변량 분석방법이라고 할 수 있다. 중도에 탈락되거나 절단된 자료와 같은 불완전한 자료에 대한 분석도 가능한데 이 경우 모형은 우도함수(likelihood function)로 표현된다.

생존함수에서 생존시간은 양의 수($t \geq 0$)라는 기본 전제 하에 t시점까지 생존할 확률을 의미하는데 생존시간 t의 확률밀도함수 $f(t)$와 누적분포 함수 $F(t)$를 다음 식과 같이 가정한다.

$$f(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t)}{\Delta t} \quad (1)$$

$$F(t) = \Pr(T \leq t) = \int_0^t f(u)du \quad (2)$$

t시점에서 생존한 사람 즉, t시점에서 연체상태를 유지하는 사람의 누적밀도 함수 F(t)라고 하면 t시점까지 연체상태를 유지할 확률은 생존 함수(S(t))

로 표현할 수 있다. 밀도함수(f(t))는 생존에서 사망으로 순간적인 확률 즉, 연체상태에서 상환상태로 바뀌는 순간적인 확률을 의미하고 생존함수의 기울기에 -1을 곱한 값이라고 할 수 있다.

$$S(t) = \Pr(T \geq t) = 1 - F(t) \tag{3}$$

$$f(t) = -\frac{dS(t)}{dt} \tag{4}$$

위험함수(h(t))는 (t-1)시점 이전에 생존에서 사망으로 전환하지 않았다는 조건 하에서 t시점에 사망할 순간적인 확률 즉, (t-1)시점 이전에 연체상태에서 상환상태로 전환하지 않았다는 조건 하에서 t시점에 연체에서 벗어나 상환할 순간적인 확률이다. 아래 식에서 보듯이 주어진 시점 때서 위험률 즉 위험함수 h(t)가 클수록 생존함수 S(t)는 작아진다.

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t+\Delta t | T \geq t)}{\Delta t} \\ &= \frac{1}{\Pr(T \geq t)} \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t+\Delta t)}{\Delta t} \\ &= \frac{f(t)}{S(t)} \end{aligned} \tag{5}$$

비례적 위험도(proportional hazard)를 이용한 준모수적(semi-parametric) Cox 모형은 우도함수를 기저위험도(baseline hazard)와 비례적 위험도로 나눈다. 여기서 기저위험도는 비모수적이며 어떤 형태이든 관계가 없는데 자료를 이용하여 추정할 수 없는 부분이다. 비례적 위험도를 나타내는 부분은 위험도를 설명하는 각 독립변수의 영향을 받고 자료를 통해 추정하는 부분이다. Cox 모형은 우도를 기저위험과 이 기저위험을 평행하게 움직이는 부분으로 나눈다. t시점에서 여러 독립변수(x_i)를 가진 상태의 위험도에 대한 함수 즉, 위험함수는 어떠한 분포도 가정하지 않는 기저위험함수(baseline hazard function)를 $h_0(t)$라 할 때 식 (6)으로 적용된다[20]. 위험함수는 확률이 아니고 단위 시간(unit of time) 당 사망률(death rate)이다. 따라서 1보다 작을 필요는 없다. 따라서 위의 식을 다시 표현하면 식 (7)과 같아진다. 이 식의 왼쪽은 결국 교차비(odds ratio)가 되고 이는 시간에 관계없이 일정하다. 즉 비례한다는 것인데 관측시점 t에서 각기 다른 독립변수의 값, 즉 x_1와 x_2를 가진 두 집단의 위험도함수 사이의 비를 구하면 초기의 비모수적 위험도 함수가 상쇄되고 독립변수에 의해 결정되는 위험도 상수의 비가 남는데 이것이 바로 비례적 위험이다(식 (8)[25]).

$$h_i(t) = \exp(\beta x_i) h_0(t) \tag{6}$$

$$\frac{h_1(t)}{h_0(t)} = \exp(\beta x_i) \tag{7}$$

$$\frac{h_1(t)}{h_2(t)} = \frac{h_0(t)\exp(\beta_1 x_1)}{h_0(t)\exp(\beta_1 x_2)} = \exp(\beta_1 (x_1 - x_2)) \tag{8}$$

$$h(t) = \frac{f(t)}{S(t)} = \frac{-\frac{dS(t)}{dt}}{S(t)} = -\frac{d}{dt}\log S(t) \tag{9}$$

여기서 이 위험비는 시간이 지나도 일정하다는 점에서 시간이 지남에 따라 위험도가 1에 가까워지는 모수적 모형과 다르다. 즉 모수적 모형에서는 시간이 지나면 위험, 즉 어떤 상태에서 벗어나는데 영향을 미치는 것은 독립변수가 아니라 시간 그 자체가 되는 것이다. 식 (5)에서는 위험함수를 생존함수와 밀도함수로 표현했으나, 위험함수를 생존함수만 식 (9)와 같이 표현할 수가 있다. 밀도함수는 생존함수와 3.1.4와 같은 관계가 있으므로 식 (5)를 다시 쓰면 식 (8)과 같다. 이와 반대로 생존함수를 위험함수만으로, 더욱 자세히는 누적위험함수(cumulative hazard function)만으로 표현할 수도 있다. 식 (9)를 다시 쓰면 식 (10)과 같으며, 여기서 $H(t) = -\int_0^t h(u)\,du$을 누적위험함수라한다.

$$\begin{aligned} &\log S(t) = -\int_0^t h(u)\,du, \text{ in other words,} \\ &S(t) = \exp[-\int_0^t h(u)\,du] = \exp[-H(t)] \end{aligned} \tag{10}$$

3.3.2 연체고객 군집별 연체탈출유형 분석 실행

본 연구에서 유형별로 분류된 연체고객의 시기별 연체탈출률과 이에 영향을 미치는 변수를 분석하기 위해 생존분석의 일종인 콕스의 비례적 위험모형을 사용한다. 일반적으로 고객신용예측에 로지스틱 회귀분석이 많이 사용되는데 이는 사건 발생 여부에 초점을 둔 방법이다. 이에 비에 콕스 모형은 생존시간에 초점을 둔 의학 분야에서 널리 이용되는 기법으로 다른 생존분석 기법과 달리 콕스 모형은 생존기간과 여러 요인들 간의 복합적인 관계를 규명하기 위해 사용된다. 사용될 자료가 백화점 신용카드 고객 중 연체고객의 자료로 이미 상환이 완료된 자료를 사용했고 부실 채권의 강제 회수로 인한 대손처리 등도 신용 데이터에서 중도 절단된 자료로 주의가 필요하지만 본 연구에서는 연체 상환이라는 사건만을 고려했기 때문에 생존분석에서 주요 이슈가 되는 탈락 혹은 절단은 문제되지 않는다.

의학 분야의 생존분석이 생존율 또는 생존함수를 통해 생존기간이 길수록 긍정적인 효과로 판단하고 영향요인을 분석하는 반면 본 연구는 연체기간에 따른 연체탈출률과 그 요인을 분석한다. 따라서 본 연구는 생존분석을 사용하지만 연체 자료 특성에 따라 생존률과 생존함수는 연체유지율과 연체유지함수로, 그리고 위험률과 위험 함수는 연체탈출률과 연체탈출 함수로 해석한다. 마지막으로 모형에서 선택된 변수들과 그 변수의 계수들로 위험함수와 생존함수를 도출하고 검증을 위해 준비된 변수 세트를 이용하여 연체탈출률이 어떻게 변하는지 알아본다. 분석을 위한 툴로 SPSS 12한글 버전의 Survival Analysis 중 Cox Regression Model을 사용한다[5].

3.4 연체고객 신용예측 시스템 적용

앞선 세 단계의 결과를 이용하여 군집별 비례적 위험모형을 생성한 후 신용예측 시스템을 구성하고 의미 있는 경영정보를 생성, 의사결정에 적용하도록 도움을 주는 단계이다. 먼저 연체고객 신용예측 시스템 구축 단계에서는 분류된 군집별 비례적 위험모형 분석의 결과를 통해 연체탈출 함수와 연체유지함수를 도출한다. 따라서 생존함수, 즉 연체유지함수에서 생존률을 t시점까지 생존할 확률로 본 연구에서는 연체기간이 지난 바로 그 다음 시점의 연체탈출률은 '1-연체유지율'로 볼 수 있다. 이런 과정으로 각 군집별 함수를 도출하여 군집별 모형을 만들고 모든 군집의 모형을 하나로 결합하면 본 연구의 목적인 연체고객 신용예측 시스템을 구축할 수가 있다. 마지막으로 검증과 비교를 거친 유의한 연체고객 신용예측 시스템을 이용하여 예상 연체탈출월과 그에 따른 예상 수입 등 경영 의사결정에 도움이 되는 경영정보를 도출하고자 한다.

4. 연체고객 신용예측 시스템 개발

4.1 연체고객 군집분류

자료의 전처리를 위해서 적합하지 않은 데이터를 제외한 41,831명의 연체 고객의 정보를 입력할 새로운 테이블을 생성하고 연산을 통해 새로운 후

〈표 4〉 군집분류를 위한 입력변수

변수명	컬럼명	비 고	변수명	컬럼명	비 고
회원번호	MEM_NO	ID	*평균연체금액*	AVG_DEL_AMOUNT	del_amount/del_time
연체횟수	DEL_TIME		*평균연체간격*	AVG_BET_PERIOD	
연체기간	DEL_PERIOD		*평균연체기간*	AVG_DEL_PERIOD	del_period/del_time
상환횟수	RETURN_TIME		*평균상환횟수*	AVG_RETURN_TIME	return_time/del_time
연체금액	DEL_AMOUNT				

보 변수들을 생성하여 <표 4>에 나타내었고 사용 여부는 *이탤릭* 폰트 변수 명으로 표시되었다. 회원 번호는 ID로 실제 분류에는 사용 되지 않는다. 변수 값의 차이가 많이 나는 입력변수들은 정규화를 거쳐 코호넨 네트워크에 입력된다.

연체고객 군집분류를 위해 SAS Enterprise Miner 6.0의 SOM/Kohonen 노드를 사용하여<Input-Replacement-SOM/Kohonen-Insight-Reporter> 모형을 구성한 후 SOM/Kohonen 노드를 Kohonen Self-Organizing Map으로 설정하고 Topological Map의 크기를 3×3로, 그리고 OM/Kohonen 노드 옵션을 설정한 후 실행한다.

실행의 결과로 <표 5>와 같은 9개의 군집이 형성 되었는데 군집 2, 3, 6과 군집 4, 5, 9는 평균연체간기간이 차이나기는 하지만, 연체 횟수, 연체금액, 상환횟수 등이 유사한 군집으로 보인다. 따라서 군집 2, 3, 6과 군집 4, 5, 9를 각각 한 군집으로 합쳐 총 5군집으로 분석을 진행한다.

<표 6>의 각 군집의 기술통계에 따르면 고객군집 A의 경우 연체횟수, 상환횟수 등으로 보아 마감일자와 상환일자 차이의 문제로 인해 연체로 등록되는 유형일 가능성이 높은 것으로 추정된다. B군집은 비록 작은 금액이라도 연 3회 정도의 많은 연체횟수를 보이는 군집으로 상습 연체고객일 가능성이 높은 군집으로 추정된다. C군집은 긴 연체기간과 분할되는 상환횟수 등을 보아 경제적으로 여의치 않다고 예상된다. D군집은 약 1회 연체를 바로 상환하는 고객군집으로 실수에 의한 연체일 가능성이 높은 것으로 추정된다. E군집은 가장 연체금액에 높은 고객군집으로 큰 금액을 연체한 후 약 2개월간, 약 1.5회에 걸쳐 상환한 고객군집이다. 개체수는 많지 않지만 다른 군집의 거의 20배에 이르는 연체금액으로 관리가 필요한 군집으로 추정된다. 본 연구는 상습 연체고객일 가능성이 있는 B군집을 중심으로 결과를 기술한다.

4.2 연체고객 군집별 연체탈출유형 분석

4.2.1 자료의 전처리 및 변수 생성

임의로 선택된 160,371명의 매출 내역 정보 중에서 연체고객 군집분류에 사용된 41,831명의 연체고객 정보와 겹치는 21,464명의 정보를 사용, 매출내역 정보와 회원 정보 요약을 위한 새로운 테이블을 생성하여 <표 7>에 나타내었다. 비례적 위험모형에서는 입력변수로 숫자만을 사용하므로 문자 정보를 숫자로 바꾸어 주고 연체탈출 여부를 표시하는 새로 생성된 STATUS 변수는 상환 기록이 있는 데이터의 값을 1로 주고 없을 경우 0으로 주었다. 물론 미상환고객의 정보는 이미 삭제한 상태의 실행한 작업으로 STATUS 값이 0인 데이터는 없

〈표 5〉 SOM/Kohonen 모형의 3×3 결과

군집	개체수	연체 횟수	평균연체 기간	평균연체 금액	평균연체 간격	평균상환 횟수	군집 재분류
1	3,735	7.60	1.19	171,310.36	1.50	1.04	A
2	5,652	4.08	1.39	151,848.84	1.96	1.15	B
3	2,149	2.09	1.44	132,145.78	7.41	1.08	B
4	1,989	1.36	6.77	192,248.32	0.52	3.91	C
5	4,931	1.54	2.90	146,435.90	0.55	2.18	C
6	6,840	2.53	1.38	132,811.57	3.44	1.13	B
7	15,543	1.09	1.08	114,572.35	0.09	1.00	D
8	143	2.21	1.81	2,807,052.06	1.19	1.48	E
9	849	1.00	10.89	236,853.37	0.00	7.88	C

〈표 6〉 각 군집의 기술통계

	개체수	구 분	연체횟수	평균연체기간	평균연체금액	평균연체간격	평균상환횟수
A군집	3,735	평균	7.60	1.19	171,310.36	1.50	1.04
		표준편차	1.67	0.23	134,142.36	0.34	0.20
		최대	12.00	2.00	1,311,837.70	2.20	1.80
		최소	6.00	1.00	8,680.00	1.00	0.40
B군집	14,641	평균	3.06	1.39	139,763.09	3.45	1.13
		표준편차	1.05	0.56	138,516.95	2.02	0.45
		최대	6.00	6.00	1,329,524.00	11.00	4.00
		최소	2.00	1.00	2.00	1.00	0.30
C군집	7,769	평균	1.43	4.77	167,947.96	0.48	3.25
		표준편차	0.68	3.12	167,045.29	0.82	2.04
		최대	4.00	12.00	2,146,900.00	6.00	11.00
		최소	1.00	1.50	2.00	0.00	0.50
D군집	15,543	평균	1.09	1.08	114,464.04	0.09	1.00
		표준편차	0.29	0.26	144,009.25	0.29	0.09
		최대	2.00	3.00	1,295,695.00	1.00	1.50
		최소	1.00	1.00	1.00	0.00	0.50
E군집	143	평균	2.24	1.83	2,753,827.22	1.23	1.48
		표준편차	1.68	1.87	2,990,921.16	1.83	1.25
		최대	8.00	11.00	20,284,134.50	9.00	10.00
		최소	1.00	1.00	1,300,156.00	0.00	0.50

다. *이탤릭* 폰트 변수명은 변수의 사용을 의미한다.

다음으로 연체고객 군집분류에서 생성된 5개의 변수와 함께 총 18개의 변수를 이용하여 변수세트를 만든다. 21,464명에 대한 변수세트를 만들고, 모형 생성과 검증을 위해 약 3 : 1의 비율로 나누어 모형입력변수세트와 검증변수세트를 만든다. 생성된 입력 변수세트는 다음 단계인 비례적 위험모형 생성단계에서 입력변수로 사용되고 검증 변수세트는 신용예측시스템 모형 검증 단계에서 모형의 결과를 검증하기 위해 사용된다.

4.2.2 연체 탈출유형 분석을 위한 비례적 위험 모형 생성

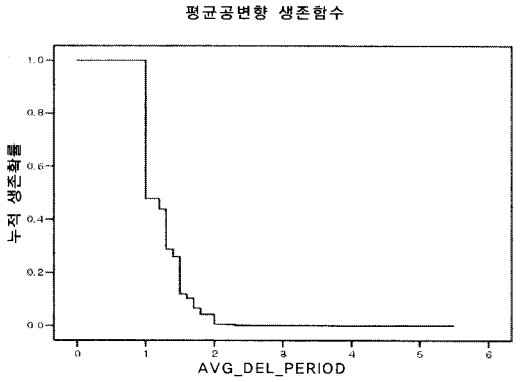
SPSS의 Survival Analysis 중 Cox Regression Model을 선택하여 사용한다. 앞선 단계에서 분류된 5개의 군집별로 따로 분석을 진행한다. 변수의 입력에 있어, 시간 변수로는 연체기간을 선택하고 상태 변수로는 STATUS를 선택한다. 나머지 변수들을 모두 입력변수로 선택하고, 범주형 변수(성별, 결혼여부 등)는 범주형으로 지정한다. "Likelihood-ratio statistics based on the conditional parameter estimate(Conditional)"로 변수 제거를 하고 "Stepwise Forward"로 변수 투입을 하였다. 모델의 유의 수준은 0.005로 설정한다. 그 결과 생성된 모형의 결과를 B군집에 대해 살펴보면 다음과 같다.

최종모형은 7단계에서 중지되었으며 '-2 Log 우도' = 106802.261, 카이제곱 = 2548.868, 자유도 = 7, 유의확률 = 0.0001로서 유의하다. 유의한 변수로 <표

〈표 7〉 연체탈출유형 분석을 위한 입력변수

변수명	컬럼명	비 고	변수명	컬럼명	비 고
구 분	GB		*결혼기념일*	MARRI_YMD	
군 집	CLUSTER_NO		*결혼여부*	MARRY	미혼/기혼(0/1)
회원번호	CUST_NO		거주시도	SIDO_NM	
매출월수	PUR_YM_CN		거주시도구분	SIDO	대구/그 외(1/2)
매출일수	PUR_YMD_CN		거주구군	SIGUN_NM	
총 매출금액	PUR_AMT_SUM		*거주구군구분*	SIGUN	
월평균 매출금액	AVG_MONTH_AMT	AVG_MONTH_AMT/ PUR_YM_CN	회원가입일	MEM_REGI_YMD	
연체 3개월 전 매출금액	AMT_3M		*회원가입기간*	MEM_REGI_PERIOD	
연체 2개월 전 매출금액	AMT_2M		카드발급일	CARD_ISSUE_YMD	
연체 1개월 전 매출금액	AMT_1M		*카드보유기간*	CARD_ISSUE_PERIOD	
연체전 3개월 평균 매출금액	AVG_3M_AMT	(AMT_3M+AMT_2M +AMT_1M)의 평균	*직업구분*	JOB_GB	전문직/회사원/ 기타(1/2/3)
나 이	AGE		직업명	JOB_NM	
성 별	SEX	남/여(1/2)	*상태*	STATUS	연체탈출(1)

8>의 변수가 선택되었고 회귀계수의 유의확률은 0.0001과 0.001로 모두 유의하다. PUR_YM_CN의 추정값(β)은 양(+)으로 매출월수가 증가할수록 위험도, 즉 연체탈출률이 증가한다. 반면 DEL_TIME, AVG_BET_PERIOD와 AVG_RETURN_TIME의 추정값(β)은 모두 음(-)으로 연체횟수, 평균연체간격과 평균상환횟수가 증가할수록 연체탈출률은 감소한다. 그리고 AVG_MONTH_AMT, AVG_3M_AMT와 AVG_DEL_AMOUNT의 추정값(β)은 0으로 평균월매출금액, 연체전 3개월 평균매출금액과 평균연체금액은 연체와 무관하다. Exp(β)로 해석하면, 매출월수, 연체 횟수, 평균연체간격과 평균상환횟수가 각각 한 단위 증가하면 연체탈출률은 각각 1.105배, 0.956배, 0.950배, 0.242배 증가한다.



[그림 2] B군집 입력변수의 평균에서 생존함수의 그래프

B군집 입력변수의 평균에서 생존함수의 그래프를 살펴보면, 연체유지율은 1개월에 50% 정도이고,

〈표 8〉 B군집 최종모형의 변수

변 수	β	유의확률	Exp(β)	변 수	β	유의확률	Exp(β)
PUR_YM_CN	0.100	0.000	1.105	AVG_DEL_AMOUNT	0.000	0.000	1.000
AVG_MONTH_AMT	0.000	0.000	1.000	AVG_BET_PERIOD	-0.051	0.000	0.950
AVG_3M_AMT	0.000	0.000	1.000	AVG_RETURN_TIME	-1.419	0.000	0.242
DEL_TIME	-0.045	0.001	0.956				

〈표 9〉 B군집 입력변수와 검증변수의 평균

구분	고객수	매출월수	연체횟수	평균연체간격	평균상환횟수	연체기간
모형	6,803	6.8	3.1	3.5	1.1	1.3
검증	2,288	6.8	3.1	3.4	1.1	1.4

2개월에 거의 0%가 되고, 3개월 내에 0%가 되는 것을 알 수 있다.

각 군집별 연체탈출률과 이에 영향을 미치는 변수간의 관계를 요약해 보면 군집별로 선택된 변수의 차이가 있기는 하지만 PUR_YM_CN(매출월수)와 AVG_RETURN_TIME(평균 상환 횟수)가 가장 중요한 변수로 사용되었다. 매출월수는 2003년 12개월 간 매출기록이 있는 월의 빈도를 나타내는 수이고 평균 상환 횟수는 각각의 연체건별로 몇 번의 상환이 있었는지를 나타내는 수로서, 이 매출월수가 증가할수록 연체탈출률이 증가하고 평균 상환 횟수가 증가할수록 연체탈출률은 감소하는 것으로 나타난다.

4.3 연체고객 신용예측 시스템 모형 검증

4.3.1 군집별 모형 검증

본 절에서는 앞서 콕스의 비례적 위험모형을 통

〈표 10〉 B군집 변수의 평균에서 모형결과와 검증결과 비교

시간(월)	기준선 누적위험함수 $H_0(t)$	누적위험함수 $H_2(t)$		생존함수 $S_2(t)$	
		모형	검증	모형	검증
1	2.430	0.739	0.734	0.478	0.480
1.3	4.104	1.248	1.240	0.287	0.289
1.4	4.442	1.351	1.342	0.259	0.261
1.5	6.979	2.122	2.109	0.120	0.121
1.6	7.473	2.272	2.258	0.103	0.105
1.7	8.901	2.707	2.690	0.067	0.068
1.8	10.347	3.146	3.127	0.043	0.044
2	17.190	5.227	5.194	0.005	0.006
2.3	21.476	6.530	6.489	0.001	0.002
2.4	22.223	6.757	6.715	0.001	0.001
2.5	29.820	9.067	9.011	0.000	0.000
2.7	32.699	9.943	9.881	0.000	0.000
2.8	34.997	10.641	10.575	0.000	0.000
3	49.932	15.183	15.088	0.000	0.000
3.5	65.026	19.772	19.649	0.000	0.000
3.7	67.126	20.411	20.284	0.000	0.000
4	94.686	28.791	28.611	0.000	0.000
4.5	102.776	31.251	31.056	0.000	0.000
5	121.209	36.855	36.626	0.000	0.000
5.5	·	·	·	0.000	0.000

한 연체고객의 연체탈출유형 분석의 결과를 이용하여 B군집에 대해 신용예측 시스템을 위한 모형을 생성하고 모형이 어느 정도의 정확도를 가지는지 검증을 하고자 한다.

앞서 생성된 B군집의 비례적 위험모형으로 누적위험함수($H_2(t) = H_0(t) \times \exp(0.1 \times$ 매출월수 $- 0.045 \times$ 연체횟수 $- 0.051 \times$ 평균연체간격 $) - 1.419 \times$ 평균상환횟수) 와 생존함수($S_2(t) = \exp[-H_2(t)]$)를 도출한다. <표 9>은 B군집의 모형에서 선택된 입력 변수와 검증 변수의 평균을 나타낸다. <표 10>은 B군집 모형입력변수와 검증변수의 평균에서 모형의 결과와 검증의 결과를 비교한 표이다.

<표 10>을 살펴보면 약간의 차이는 있지만 변수의 평균에서 누적위험 함수와 생존함수의 각 시점별 변화량 및 변화추세는 거의 비슷한 것을 볼 수 있다. <표 11>은 B군집 검증 변수세트를 앞서 생성된 생존함수에 입력한 결과를 각 회원별로 나누어 표로 작성하고 평균으로 정리한 것이다.

생성된 모형의 각 군집별 실제탈출월에서의 예측된 탈출 확률을 종합해 평균으로 정리해 보면 다음과 같다.

모형 전체의 예측탈출률의 평균이 97.5%로 실제 탈출월의 평균 2.3개월에서 연체를 탈출할 확률이 약 97%인 것으로 나타난다. 즉 모형의 입력 변수의 실제 연체탈출월에서의 실제 연체탈출률을 100%로 본다면 검증 변수의 예측탈출률은 97.5%로 2.5%의 차이가 난다는 것으로 모형 전체의 예측력도 매우 정확하다고 볼 수 있다. 또한 C군집에서 보듯 전체적으로도 연체기간이 길어질수록 예측력이 조금씩 떨어지는 모습을 보여 준다.

4.4 로지스틱 회귀분석과 비교

비례적 위험모형의 결과의 상대 비교를 위해 각 군집에 대한 로지스틱 회귀분석 모형과 비례적 위험모형에 검증 변수 입력 결과의 평균을 이용하여 다음과 같은 표를 작성하였다. 로지스틱 회귀분석 모형은 주로 불량채권 발생 유/무를 예측하는 등 한 시점에서의 이벤트 즉, 사건의 발생에 영향을 주는 인자의 선택과 그 사건이 발생할 확률을 산출하기 위해 사용되지만, 여기서는 앞서 생성한 비례적 위험 모형과의 비교를 위해 각 기간별 연체상태 유지를 기준으로 연체탈출확률을 구하는 모형을 생성한 다음 두 모형의 결과를 비교하였다.

〈표 11〉 B군집 검증결과

회원번호	실제연체기간	실제탈출월	예측탈출률	예측연체유지율				
				1개월	2개월	3개월	4개월	5개월
A00000039	1	2	100.0%	0.236	0.000	0.000	0.000	0.000
A00000211	1	2	100.0%	0.234	0.000	0.000	0.000	0.000
A00000586	1.7	2	82.0%	0.785	0.180	0.007	0.000	0.000
A00000757	1.5	2	83.0%	0.779	0.170	0.006	0.000	0.000
A00001028	1.2	2	96.9%	0.611	0.031	0.000	0.000	0.000
A00001091	1.4	2	100.0%	0.340	0.000	0.000	0.000	0.000
A00001140	1.7	2	93.6%	0.678	0.064	0.000	0.000	0.000
A00001247	2	3	94.7%	0.867	0.364	0.053	0.004	0.004
A00001682	1	2	99.8%	0.411	0.002	0.000	0.000	0.000
A00001849	1	2	100.0%	0.152	0.000	0.000	0.000	0.000
…	…	…	…	…	…	…	…	…
평균	1.4	2.2	98.0%	0.453	0.052	0.010	0.004	0.004

〈표 12〉 군집별 검증결과 종합

구 분	실제연체기간	실제탈출월	예측탈출률
A군집	1.2	2.0	100.0%
B군집	1.4	2.2	98.0%
C군집	3.5	4.4	84.1%
D군집	1.1	2.0	100.0%
E군집	1.7	2.6	97.1%
평 균	1.4	2.3	97.5%

〈표 13〉 군집별 비례적 위험모형과 로지스틱 회귀분석의 결과 비교

구 분	실제연체 기간	실제 탈출월	예측탈출률	
			비례적 위험모형	로지스틱 회귀분석
A군집	1.2	2.0	100.0%	99.4%
B군집	1.4	2.2	98.0%	93.4%
C군집	3.5	4.4	84.1%	71.3%
D군집	1.1	2.0	100.0%	94.8%
E군집	1.9	2.8	97.1%	90.7%
전체	1.4	2.3	97.5%	92.3%

<표 13>을 살펴보면 군집 전체에서 실제탈출월의 평균 2.3개월에서 로지스틱 회귀분석 모형과 비례적 위험모형의 예측탈출률의 평균은 각각 92.3%와 97.5%로 비례적 위험모형의 결과가 5.2% 가량 높다. 각 군집별로도 그 차이는 0.6%, 4.6%, 12.8%, 5.2% 그리고 6.4%로 모든 군집에서 비례적 위험모형의 결과가 높았음을 알 수 있다.

4.5 연체고객 신용예측 시스템에 의해 예측된 경영정보

본 절에서는 앞서의 연체고객 신용예측 시스템의 예측결과를 통해 각 고객의 연체탈출 후 예상 수입을 계산해 보고 분석해 보고자 한다. 연체탈출예측률이 100%가 되는 월을 연체예측탈출월로, 평균연체금액을 연체탈출 후 환수되어질 연체금액으로, 월별평균매출을 예상되는 월매출로 정의하여 예측탈출월에 환수되는 연체금액과 예측되는 월매출을 합해진 예상 수입을 각 고객별로 따로 계산하여 그 합계, 평균과 빈도 등으로 정리한다. 각 예측탈출월별로 예상탈출빈도와 백분율을 군집별로 계산해 <표 14>에 나타내었는데 2개월에서의 예상탈출빈도가 74%로 가장 많다.

다음으로 각 예측탈출월별로 예상수입합계와 백분

〈표 14〉 예측탈출 월별 각 군집의 예상탈출빈도와 백분율

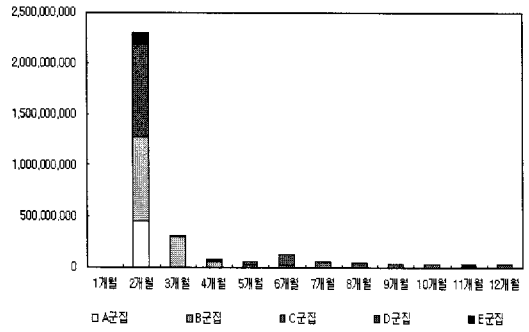
구 분	예상 연체탈출 빈도						
	A군집	B군집	C군집	D군집	E군집	합계	백분율
1개월	0	0	0	0	1	1	0%
2개월	620	1,460	0	1,889	18	3,987	74%
3개월	0	637	2	0	2	641	12%
4개월	0	124	15	0	1	140	3%
5개월	0	25	61	0	0	86	2%
6개월	0	42	190	0	0	232	4%
7개월	0	0	95	0	1	96	2%
8개월	0	0	70	0	0	70	1%
9개월	0	0	43	0	0	43	1%
10개월	0	0	36	0	0	36	1%
11개월	0	0	17	0	0	17	0%
12개월	0	0	16	0	0	16	0%
합계	620	2,288	545	1,889	23	5,365	100%

〈표 15〉 예측탈출 월별 각 군집의 예상수입합계와 백분율

구 분	예상 수입(단위 : 천원)						
	A군집	B군집	C군집	D군집	E군집	합 계	백분율
1개월	0	0	0	0	4,311	4,311	0%
2개월	441,051	830,632	0	931,877	97,560	2,301,119	74%
3개월	0	299,373	1,693	0	8,107	309,173	10%
4개월	0	60,370	6,688	0	8,423	75,480	2%
5개월	0	17,057	41,435	0	0	58,491	2%
6개월	0	24,779	98,244	0	0	123,023	4%
7개월	0	0	50,636	0	0	50,636	2%
8개월	0	0	47,111	0	0	47,111	2%
9개월	0	0	31,967	0	0	31,967	1%
10개월	0	0	38,606	0	0	38,606	1%
11개월	0	0	15,623	0	12,428	28,051	1%
12개월	0	0	35,982	0	0	35,982	1%
합계	441,051	1,232,210	367,984	931,877	130,829	3,103,951	100%

율을 군집별로 계산해 표로 정리하면 <표 15>와 같다. 이 표를 살펴보면 2개월에서 예상수입금액이 74%로 가장 많고, 그 다음이 3개월로 10%이다.

더 직관적인 분석을 위해서 예측탈출월별 탈출 후 예상수입에 대한 그래프를 작성하면 [그림 3]과 같다.



[그림 3] 각 군집별 예측탈출월별 탈출 후 예상 수입

예상수익 그래프 또한 색상을 통해 군집을 구분한다. 그래프를 살펴보면 연체 후 2개월에서 연체에서 탈출하여 회수되고 신규로 발생될 매출이 가장 높다. 두 그래프를 살펴보면 역시 2개월에서 가장 높고 그 다음으로 3개월이 조금 높지만 아주 많지는 않으며 나머지는 낮고 유사한 추세를 보여준다. 특히 군집 A, B, D군집의 경우 대부분 연체 후 2개월에서 연체 탈출이 이루어지고 예상수입이 발생한다는 것을 알 수 있는데 2개월에 발생할 예상수입에 대한 관리가 필요할 것으로 보인다. 특히 군집 A의 경우는 마감일자와 상환일자에 대한 고객의 착각으로 인해 연체자로 분류될 수 있는 가능성이 큰 군집이기 때문에 결제일 자체의 조정을 권고하는 이메일을 보내거나 결제일을 미리 알려주는 SMS 등을 보내는 것, 또는 자동 이체를 권유하는 방법 등이 연체를 사전에 막는 방법이 될 수 있을 것이다.

B군집의 경우 상습연체 고객 군집으로 분류되는데 이의 예방을 위해 결제일 안내 SMS나 이메일을 통해 상환을 독려하고 연체가 계속될 경우 신용한도 조정 등의 작업이 필요할 것이다. C군집의 경우 경제적 여건이 좋지 않은 군집으로 예상되었기 때문에 신용 한도 조정을 통해서 적절한 구매 한도를 정하는 것이 연체를 최소화 할 수 있을 것이다. D군집의 경우 고객의 결제일 착각으로 인해 연체가 되는 경우로 예상되므로 자동 이체의 독려, 결

제일 안내 시스템 등이 고객 관리에 도움이 될 것으로 보인다.

<표 16>은 각 군집별 예상수입의 차이를 살펴보기위해, 각 군집별로 예측탈출월에서의 환수되는 연체금액, 예상되는 월매출 및 그 합계의 평균과 총합을 계산해 나타내었다. 이 표에서 예상 수입은 연체 이전 평균 월매출을 근거로 예측되어진 것이다.

<표 16>을 살펴보면, E군집의 빈도는 가장 낮은 반면에 각 고객별 평균 금액은 매우 높은 것을 알 수 있다. 이를 다시 살펴보기 위해 군집별 평균 연체금액과 월별평균매출로 구분한 탈출 후 예상수입의 평균에 대한 그래프를 작성하면 [그림 3]과 같고 이를 살펴보면, 역시 앞서 예상된 것처럼 E군집이 다른 군집에 비해 굉장히 큰 평균 예상 수입을 보여 준다. 즉 연체탈출 후 회수될 금액과 예상되는 매출 모두가 다른 군집에 비해 굉장히 큼을 알 수 있다. 그러나 예상 매출액이 가장 큰 E군집의 고객은 위험 관리의 측면에서 연체 금액 또한 크기 때문에 연체 금액의 회수가 어렵고 반복적으로 연체가 이루어진다면 고액 연체에 따른 위험 고객으로 분류하여 총 신용 금액 등의 조정 작업이 필요한 것을 암시한다고도 볼 수 있다.

5. 결 론

5.1 연구의 요약 및 의의

본 연구는 국내 백화점의 공통적인 문제점인 백화점 신용카드 연체고객의 관리방안의 일환으로 D백화점의 신용카드 연체기록이 있는 고객의 정보를 유형별로 군집을 만들고 각 군집별 연체탈출 유형을 분석하여 연체고객관리에 도움을 줄 수 있는 연체고객신용예측 모형을 제시하였다.

본 연구 모형은 연체고객 분류를 위해 코호넨 네트워크와 비례적 위험모형을 사용하고 군집별로 연체탈출의 예상기간과 그 연체탈출에 영향을 주는 변수들을 선정하여 고객의 신용예측에 적용할 수 있도록 했다. 본 연구에서는 연체의 유형 즉 연체횟수, 연체기간, 연체금액, 연체간격, 상환횟수 별로 연체고객이 5개의 군집으로 나누어 졌다. 다음으로 앞서 분류된 군집별로 연체 탈출 유형을 분석해 기간별 연체탈출률과 각 변수가 연체탈출률에 미치는 영향을 살펴보았다. 군집별로 선택된 변수의 차이가 있기는 했지만 매출월수와 평균 상환 횟수가 가장 중요한 변수로 사용되었고 이것은 매출월수가 증가할수록 연체탈출률이 증가하고 평균 상환 횟수가 증가할수록 연체탈출률이 감소한다는 것을 의미한다. 앞선 과정의 결과를 통해 만들어진 신용예측 시스템은 전체적으로 97.5%의 예측력을 보여 주었다.

본 연구의 의의는 첫째, 본 연구는 기존 신용예측시스템의 우량·불량고객으로 구분되는 이분법적 모형이 아닌 생존분석을 이용하여 연체탈출에 영향을 주는 변수를 선택하고 그 변수가 주는 영향의 크기를 도출할 뿐만 아니라 연체탈출예상기간과 예측탈출률을 계산할 수 있게 해준다는 것이다. 둘째, 본 연구는 기존의 이분법적 모형에서는 논의되

〈표 16〉 군집별 예상수입의 평균과 총합

구 분	환수되는 연체금액		예상되는 월매출금액		금액합계	
	평 균	총 합	평 균	총 합	평 균	총 합
A군집	176,199	109,243,371	535,173	331,807,237	711,372	441,050,608
B군집	144,182	329,889,234	394,371	902,321,029	538,553	1,232,210,263
C군집	166,295	90,630,556	508,906	277,353,644	675,200	367,984,200
D군집	132,267	249,852,891	361,050	682,023,886	493,318	931,876,777
E군집	2,410,088	55,432,024	3,278,135	75,397,112	5,688,223	130,829,136
전체	155,647	835,048,076	422,908	2,268,902,908	578,556	3,103,950,984

지 않았던 회복가능 고객 군집에 대한 분석을 가능하게 해준다는 것이다. 셋째, 본 연구는 단순히 부실가능성 측정이 아닌 부실로 인한 손실의 추정과 수익에 미치는 영향까지 추정할 수 있는 모형이다. 즉 연체고객 신용예측 시스템을 통해 제공될 세분화된 고객군집의 기간별 회수가능 불량채권에 대한 정량적 정보를 제공하여 백화점 신용판매 관리부서나 다른 경영지원 부서에서도 전략적으로 중요한 정보로 사용될 수 있게 해준다는 것이다. 물론 고객신용 예측모형이 연체고객관리 전략 등의 수립 시 활용 또는 의사결정을 위한 참고자료일 뿐, 그 자체가 절대 해답이 될 수는 없겠지만, 최종적으로 모형을 통해 얻어진 예상 연체상환금을 예상 매출로 전환하여 사업계획 수립, 기업자원 배분 등 다양한 경영의사결정에 적용하여 백화점의 생존을 위한 든든한 의사결정지원시스템의 일환으로 자리 잡을 수 있을 것으로 기대된다.

5.2 연구의 한계점 및 향후 과제

본 연구는 다음과 같은 몇 가지 한계점을 가지고 있으며 이는 향후 연구에서 개선되어야 할 사항들이다. 첫째, 데이터 수집상의 어려움으로 인해 D백화점의 2003년 12개월 간의 연체 고객에 대한 자료만을 사용하였기 때문에 연구 결과의 일반화를 위해 다른 기간의 데이터를 확보하여 분석할 필요가 있다. 둘째, 본 연구에서 사용된 변수들은 백화점 관리부서의 도움으로 연체관리에 중요하다고 판단되어 도출된 변수들이다. 하지만 본 연구에서 고려되지 않은 새로운 파생변수가 생성될 수 있을 가능성에 대해서도 추후에 연구가 필요하다.

셋째, 본 연구에서는 군집 분류 기법으로 코호넨 네트워크를 사용하고 분류된 군집에서 상환률을 예측하기 위한 기법으로 비례적 위험모형을 사용하였다. 논문의 예측률이 95%를 상회하는 높은 성과를 보이고는 있지만 모집된 데이터의 한계로 인해 장기간 연체율이 유지되는 데이터의 경우 새로운 모형을 구성하는 등 계속적인 예측 모형 개발이 필요하다.

넷째, 본 연구는 신용회복가능 군집에 대한 분석을 진행하였지만 추후에 우량고객과 불량고객(채권미회수) 군집에 대한 분석까지 모두 포함한 전체 고객에 대한 신용예측 시스템으로 확대하여 연구할 필요가 있다.

참 고 문 헌

[1] 김갑식, "신용평가를 위한 데이터마이닝 분류모형의 통합모형에 관한 연구", 「정보처리학회논문지」, 제12권, 제2호(2005), pp.211-218.

[2] 김갑식, 이동만, 황하진, "유전자알고리즘을 이용한 할부금융회사의 고객 신용평가 데이터마이닝 모형 구축", 「경영연구」, 제3권, 제4호(2003), pp.249-272.

[3] 김대수, 「신경망 이론과 응용」, 하이테크 정보, 서울, 1992.

[4] 김명진, 서용무, "차별적 대응방안 수립을 위한 재발연체 고객의 분류모형", 「한국경영정보학회 추계학술대회 논문집」, 2004, pp.797-804.

[5] 송경일, 안재억, 「SPSS for Windows를 이용한 생존분석」, SPSS아카데미, 서울, 1999.

[6] 이웅규, 김홍철, "유전자 알고리즘기반 복수분류모형 통합에 의한 할부금융고객의 신용예측모형", 「한국경영과학회 추계학술대회 논문집」, 2001, pp.161-164.

[7] 이병윤, "신용위험평가 방법의 현황 및 전망", 「은행경영 브리프」, 제13권, 제17호(2004), pp.18-21.

[8] 정충영, 최이규, 「SPSSWIN을 이용한 통계분석」, 4판, 무역경영사, 서울, 2001.

[9] 최종후, 한상태, 강현철, 김은석, 김미경, 「SAS Enterprise Miner 4.0을 이용한 데이터 마이닝 : 기능과 사용법」, 자유아카데미, 서울, 2001.

[10] Allen, L.N. and L.C. Rose, "Financial survival analysis of defaulted debtors," *Journal of the Operational Research Society*, Vol.57, No.6(2006), pp.630-636.

[11] Altman, E.I., G. Marco, and F. Varetto, "Corporate Distress Diagnosis : Comparisons Using Linear Discriminant Anlysis and Neural Networks (the Italian Experience)," *Journal of Banking and Finance*, Vol.18(1994), pp.505-520.

[12] Andreeva, G., "European Generic Scoring Models Using Survival Analysis," *CRC Working Papers*, Vol.74, No.2(2004).

[13] Baesens, B., M. Egmont-Petersen, R. Castelo, and J. Vanthienen, "Learning Bayesian Network Classifiers for Credit Scoring Using Markov Chain Monte Carlo Search," *Proceedings of the 16th International Conference on Pattern Recognition*, (2002), pp.49-52.

[14] Baesens, B., T.V. Gestel, M. Stepanova, and D.V. Poel, "Neural Network Survival Analysis for Personal Loan Data," *Journal of the Operational Research Society*, Vol.59, No.9(2005), pp.1089-1098.

[15] Berry, M.J.A. and G. Linoff, *Data Mining Techniques : For Marketing, Sales, and Customer Support*, Wiley and Sons, (2004).

[16] Bradley, P.S., U.M. Fayyad, and O.L. Mangasarian, "Data Mining : Overview and Optimization Opportunities," *INFORMS*, Special issue on Data Mining, (1998), pp. 17-22.

[17] Carter, C. and J. Catlett, "Assessing Credit Card Applications Using Machine Learning," *IEEE Expert*, Vol.2(1987), pp.71-79.

[18] Chen, M.C. and S.H. Huang, "Credit Scoring and Rejected Instances Reassigning through Evolutionary Computation Techniques," *Expert System with Applications*, Vol.24(2003), pp.433-441.

[19] Cheng, B. and D.M. Titterington, "Neural Networks : A Review from a Statistical Perspective," *Statistical Science*, Vol.9(1994), pp.2-30.

[20] Cox, D.R., "Regression Models and Life-Tables," *Journal of Royal Statistical Society*, Vol.26 (1972), pp.187-202.

[21] David, W., "Neural Network Credit Scoring Models," *Computers and Operations Reserch*, Vol.27(2000), pp.1131-1152.

[22] Desai, C.S., D.G. Convay, J.N. Crook, and G.A. Overstreet, "Credit Scoring Models in the Credit Union Environment Using Neural Networks and Genetic Algorithms," *IMA Journal of Mathematics Applied in Business and Industry*, Vol.8(1997), pp.323-346.

[23] Desai, C.S., J.N. Crook, and G.A. Overstreet, "A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment," *European Journal of Operational Research*, Vol.95(1996), pp.24-37.

[24] Grablowsky, B.J. and W.K. Talley, "Probit and Discriminant Functions for Classifying Credit Applicants : A Comparison," *Journal of Economics and Business*, Vol.33(1981), pp.254-261.

[25] Greene, W.H., *Econometric Analysis*, 3rd Edition, Prentice-Hall, Inc., 1997.

[26] Gupta, Y.P., M.C. Gupta, A.K. Kumar, and C. Sundram, "Minimizing Total Intercell and Intracell Moves In Cellular Manufacturing. A Genetic Algorithm Approach," *International Journal of Computer Integrated Manufacturing*, Vol.8, No.2(1995), pp.92-101.

[27] Han, J. and M. Kamber, *Data Mining : Concepts and Techniques*, 2nd Edition, Morgan Kaufmann Publishers, CA, (2004).

[28] Hand, D.J. and W.E. Henley, "Statistical

Classification Methods in Consumer Credit Scoring : A Review," *Journal of the Royal Statistical Society*, Vol.162(1997), pp.523-541.

[29] Hansen, J.V., "Combining Predictors : Comparison of Five Meta Machine Learning Methods," *Information Science*, Vol.119(1999), pp.91-105.

[30] Hon, K.K.B. and H. Chi, "A New Approach of Group Technology Part Families Optimization," *Annals of the CIRP*, Vol.43, No.1(1994), pp.425-428.

[31] Hsieh, N.C., "Hybrid mining approach in the design of credit scoring models," *Expert Systems with Applications*, Vol.28, No.4(2005), pp.655-665.

[32] Hu, X., "A Data Mining Approach for Retailing Bank Customer Attrition Analysis," *Applied Intelligence*, Vol.22, No.1(2005), pp.47-60.

[33] Huanga, C.L., M.C. Chenb, and C.J. Wang, "Credit scoring with a data mining approach based on support vector machines," *Expert Systems with Applications*, Vol.33, No.4(2007), pp.847-856.

[34] Imielinski, T. and H. Mannila, "A Database Perspective on Knowledge Discovery," *Communications of the ACM*, Vol.40, No.11 (1996), pp.214-225.

[35] Jain, B.A. and B.N. Nag, "Performance Evaluation of Neural Network Decision Models," *Journal of Management*, Vol.14, No.2(1997), pp.201-215.

[36] James, H.M. and W.F. Edward, "The Development of Numerical Credit Evaluation Systems," *Journal of the American Statistical Association*, Vol.58(1963), pp.799-806.

[37] Kim, E., W. Kim, and Y. Lee, "Purchase Propensity Prediction of EC Customer by Combining Multiple Classifiers Base on GA," *Proceedings of International Conference on Electronic Commerce*, (2000), pp.274-280.

[38] Kohonen, T., "The Self-Organizing Map," *Proceedings of the IEEE*, Vol.78, No.9(1990), pp.1464-1480.

[39] Lee, T.S., C.C. Chiu, C.J. Lu, and I.F. Chen, "Credit Scoring Using the Hybrid Neural Discriminant Technique," *Expert Systems with Applications*, Vol.23(2002), pp.245-254.

[40] Mangasarian, O.L., "Linear and Nonlinear Separation of Patterns by Linear Programming," *Operations Research*, Vol.13(1965), pp.444-452.

[41] Marron, D., "Lending by numbers : Credit Scoring and the Constitution of Risk within American Consumer Credit," *Economy and Society* Vol.36(2007), pp.103-133.

[42] Mehta, D., "The Formulation of Credit Policy Models," *Management Science*, Vol15(1968), pp.30-50.

[43] Sarlija, N., M. Bensic, and Z. Bohacek, "Customer Revolving Credit-How the Economic Conditions Make a Difference," *CRC 2005 Credit Scoring Conference Archive*, Vol.27(2005).

[44] Sarlija, N., M. Bensic, and M. Zekic-Susac, "A neural network classification of credit applicants in consumer credit scoring," *Proceedings of the 24th IASTED international conference on Artificial intelligence and applications*, (2006), pp.205-210.

[45] Thomas, L.C., "A Survey of Credit and Behavioral Scoring : Forecasting Financial Risk of Lending to Consumers," *International Journal of Forecasting*, Vol.16(2000), pp.149-172.

[46] Thomas, L.C., J. Ho, and W.T. Soberer,

"Time Will Tell : Behavioral Scoring and the Dynamics of Consumer Credit Assessment," *IMA Journal of Management Mathematics*, Vol.12(2001), pp.89-103.

[47] West, D., "Neural Network Credit Scoring Models," *Computers and Operations Research*, Vol.25(2000), pp.1131-1152.

[48] Wiginton, J.C., "A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behaviour," *Journal of Financial and Quantitative Analysis*, Vol.15(1980), pp.757-770.

[49] Jiao, Y., R. Syou, and E.S. Lee, "Modeling Credit Rating by Fuzzy Adaptive Network," *Mathematical and Computer Modeling*, Vol.45(2007), pp.717-731.