석 사 학 위 논 문

데이터마이닝을 활용한 신용평가모형 연구

고려대학교 정책대학원

통계조사학과

용 관

2009년 12월 18일

조 형 준 교수지도
석 사 학 위 논 문

데이터마이닝을 활용한 신용평가모형 연구

이 논문을 통계학 석사학위 논문으로 제출함.

2009년 12월

고려대학교 정책대학원
통계조사학과
용 관

용 관의 통계학석사 학위논문

심사를 완료함.

2009년 12월 일

위 원 장 (인)

위 원 (인)

위 원 (인)

국 문 요 약

신용평가를 위한 모형구현과 신용평점표 작성에 관한 일반적인 과정에 대해 연구하였다. 자료를 획득하여 모형화하고 신용평점표를 작성하는 일련의 과정을 데이터마이닝 기법에 접목하였다. 본 연구는 원자료에 대한 이해, 자료의 정제와 모형화, 신용평점표 작성까지 총 3개의 과정으로 구성되어있다.

첫 번째 과정인 원자료의 이해과정에서는 자료의 분포를 관찰하고 적절한 변환과 이산화과정을 통해 분석하기 용이한 새로운 자료를 구축하였다.

두 번째는 구체적인 모형화 과정으로 의사결정나무분석과 회귀분석 두 가지 방법에 대해 연구하였다. 의사결정나무분석에서는 세가지 알고리즘을 토대로 모형을 구축하였고 여러 가지 평가기준을 바탕으로 최적의 모형을 도출하였다. 의사결정나무분석은 선택된 최종모형에서 어떠한 변수들이 신용상태에 영향을 미치는지 1차적으로 보여준다. 회귀분석 방법에서는 로지스틱 회귀분석 방법을 사용하였고 세 가지 변수선택방법에 따른 각각의 모형에 대해 Lift Graph, ROC curve, 각종 검정통계량을 기준으로 최적의 모형을 선별하였다. 의사결정나무분석의 최적모형과 로지스틱 회귀분석의 최적모형을 비교하여 신용평점표산출을 위한 최종모형을 선택하였고 Lift Graph와 ROC curve를 기준으로 평가하여 로지스틱 회귀분석 방법에 의한 모형을 최종 선정하였다.

세 번째는 신용평점표를 작성하는 과정이다. 신용평점표 작성을 위해 자료의 정제과정에서 범주화된 변수들을 재범주화하였고 통계적으로 유의한 각 동일 변수 내부의 범주별로 가중치를 부여하여 평점화하여 신용평점표를 작성하였다. 산출된 신용평점표를 실제로 활용가치가 있는지 확인하기 위해 K-S통계량과 ROC curve의 C-통계량을 기준으로 타당성을 평가하였고 PSI를 기준으로 개발용 데이터와 검증용 데이터를 비교하여 변동여부 및 안정성을 평가하였다.

결측치에 대한 처리와 상관분석, 군집분석등 고차원적인 방법으로 데이터를 선별하거나 신경망분석, 복합방법론등 다양한 방법을 활용하여 모형을 구현하지는 못하였으나 본 연구를 통해 자료 정제의 중요성을 확인할 수 있었고 신용평가를 위한 모형구축과 신용평점표 산출의 전반적인 과정을 이해할 수 있었다는 것이 본 연구의 가장 큰 성과라 할 수 있다.

목 차

표 목 차

그 림 목 차

제 1 장 서 론

1.1 연구배경

양화가 악화를 구축한다는 그리샴의 법칙처럼 화폐경제보다 신용의 개념이 더욱 발전하여 금융기관의 전통적인 역할인 수신과 여신의 비중이 작아지면서 운용의 묘와 수익이 중시되고 있다. 2008년 세계 경제를 세계 대공황이후 최대의 위기로 몰아간 가장 큰 원인도 금융의 급속한 발전에 따른 부작용이라할 수 있다. 자금의 흐름을 원활하게 하고 소액의 자본금으로 수십배 이상의 자본을 운용할 수 있는 레버리지 효과는 빠른 성장과 다수의 경제주체들에게 매력적인 금융 수익원으로 자리매김하였지만 뿌리를 알 수 없는 파생상품과 리스크 관리에 소홀했던 금융정책은 지금까지 다져온 경제기반을 모조리 흔들 수 위험요인이 되고 있다.

세계의 경제가 건실하기 위해서는 각 국가들의 경제기반이 탄탄해야하며 한 국가의 경제가 굳건하기 위해서는 경제주체의 최소 단위인 개인의 경제상태가 건전해야한다. 개인의 재무상태가 건전하다면 국가 경제는 자연스럽게 튼튼해질 수 있다. 금융시장에서 개인의 역할은 중요하다. 한 개인의 부실은 전체적인 입장에서 본다면 작은 상처에 불과하지만 작은 상처들이 곪고 퍼져 치료할 수 없는 지경의 큰 병이 될 수 있는 것처럼 다수의 개인부실은 국가전체에 치명적인 부실원인이 될 수 있다. 2003년 카드대란 이후 2008년 세계 금융위기가 발생하기까지 5년 동안 우리나라 신용불량자의 수와 그 비중은 꾸준한 증가추세에 있었고 언제 폭발할지 모르는 시한폭탄이다. 또한 아르헨티나, 아일랜드처럼 국가부도의 도화선이 될 수 있는 위험요인이기도 하다. 개개의 경제주체 재무건전도는 국가경제 전체 건전도의 충분조건이 된다고 하여도 과언이 아니다.

2009년 6월말 현재 우리나라 개인 1인당 부채가 1678만원을 넘어섰으며 국가 전체적으로는 818조 4천억원에 육박[1]하고 있다. 이는 국가경제의 분명한 적신호이다. 금융기관들은 이러한 위험신호에 대비하여 부실 대출을 방지하기

1) 한국은행 2/4분기 자금순환동향 (한국은행, 2009)

위해 갖은 노력을 기울이고 있다.

과거에는 금융기관 상호간 고객정보를 공유하지 못해 체계적으로 위험을 관리하지 못하였으나 현재에는 은행연합회를 중심으로 정보를 통합관리하여 개인신용정보를 효율적으로 관리함으로써 부실을 최소화하기 위해 공조하고 있다. 은행을 중심으로한 개인신용평가관리시스템은 각 은행별로 특화되어 독자적으로 부실을 방지하고 있으나 우리나라 8개 시중은행이 관리하는 ARM(Analysis Risk Management) system은 과거 고객들의 거래실적을 토대로 위험확률을 예측하는 공통적인 개념적 방법을 취하고 있다. 신용우량자와 불량자의 고객정보와 거래패턴을 중심으로 우량, 불량확률을 예측하고 우량집단과 불량집단으로 분류될 확률을 평점화된 모형을 바탕으로 개인에게 신용등급을 부여하는 방법을 선택하고 있는 것이다.

신용등급은 단순한 상태나 지표에서 벗어나 경제적, 사회적 계급으로 자리잡아가고 있다. 금융기관이 개인의 신용도에 따라 고객을 구분하는 신용차별화현상은 금융기관들이 스스로의 건전성을 높이고 부실을 방지하기 위한 노력의 결과이며 각 기관과 회사들은 개인의 신용상태를 공유하며 체계적인 시스템을 꾸준히 발전시켜왔다. 개인과 금융회사 모두 공감하고 사회적으로 인정받을 수 있는 제도가 정착되기 위해서는 개인의 경제활동 정보를 토대로 체계적인 시스템을 구축하여 객관적으로 관리하여야 한다. 실증적인 방법의 한 형태가 신용평점제도이다. 신용평점에 대한 금융기관의 의존도는 갈수록 높아지고 있다. 본론에서는 일반 시중은행에서 공통적으로 채택하고 있는 신용평점 모형을 도출하는 과정을 연구하였으며 데이터마이닝기법중 의사결정나무, 회귀분석의 방법을 비교분석하여 개인신용평점표 산출과정으로 연결되는 일련의 절차를 이해하고자 하였다.

1.2 연구목적

본 연구의 목적은 두 가지이다. 첫 번째는 변수변환의 중요성을 밝히는 것이고 두 번째는 신용평가예측모형구축과 신용평점표 작성에 대한 전반적인 과정을 이해하는 것이다. 각각의 변수들의 특성을 파악하고 변수변환 과정을 통해 변환된 자료들이 원자료에 비하여 활용가치가 높다는 것을 확인하고 각 방법론의 이론적인 요소와 실무경험을 바탕으로 직관적인 경험적 요소도 가미하여 변수를 식별하면서

적절한 모형에 접근해가는 과정을 탐색하는데 있다. 의사결정나무와 회귀분석의 각 방법론의 내부적으로는 알고리즘과 변수선택방법을 비교하면서 최적의 모형을 선별하는 과정을 탐색하는 것과 외부적으로는 두 방법론의 예측성능을 비교하여 신용평점표모형을 완성하기 위한 연결과정을 연구하였다.

1.3 분석방법 및 연구절차

원자료 정제의 첫 번째 과정은 결측치를 찾아내고 제거하여 정확성을 향상시키는데 중점을 두었다. 결측치 제거 전후의 결과를 비교하여 결측치 제거과정의 중요성을 제시하였다.

변수변환 과정에서는 변수를 적절한 형태로 변환시켜 유의하지 않은 변수가 모형을 구축하는 과정에서 유용하게 활용될 수 있도록 유도하였다. 각 변수들의 기초통계량과 분포를 확인하여 연속형 변수를 순서형, 명목형으로 변환하기 위한 기준을 설정하였고 변수들의 일부 정보가 손실되더라도 더욱 가치있는 변수로 활용될 수 있다는 사실을 확인하였다.

연구의 기본 도구로써 SAS E-miner를 사용하였으며 데이터마이닝의 대표적 방법인 의사결정나무, 회귀분석을 적용하였다. 종속변수를 신용상태인 우량, 불량으로 정의하고 비모수적 방법인 의사결정나무를 바탕으로 신용상태에 영향을 주는 주요변수들을 식별하였다. 회귀분석방법에서는 각 독립 변수들이 얼마나 신용상태에 영향을 주고 있는지 확인하는데 중점을 두었다. 의사결정나무와 회귀분석방법을 비교하여 모형의 예측성능을 비교하고 평점표작성에 활용하는 이유를 제시하였다.

의사결정나무와 회귀분석방법중 예측성능이 상대적으로 우수한 회귀분석방법을 평점표 작성 방법으로 선택하였다. 회귀분석방법은 의사결정나무에 비해 안정성, 해석의 용이성과 실무에서의 활용성등을 종합적으로 고려해볼 때 의사결정나무 방법보다 효율적이기 때문에 평점표산출 최종모형으로 회귀분석방법을 선택하였고 선별된 변수들을 다시 범주화하여 수개의 구간으로 분할하고 유의한 변수들을 선별하였다. 선별된 변수들의 각 범주에 가중평점을 부여하여 평점표를 완성하였으며 완성된 평점표에 대한 타당성과 안정성 평가를 통해 신용평점표로서 역할을 할 수 있는지에 대해 확인하였다.

제 2 장 자료의 이해

안정된 신용평가 모델을 구축하기 위해서는 수 많은 데이터가 필요하다. 그러나 데이터의 수가 많으면 많을수록 전통적인 통계분석에서 활용하는 이상적이고 정적인 데이터의 특질과는 거리가 멀어지고 데이터의 저장과 보관, 편집, 변환과 같은 사소한 부분을 포함하여 결측치 발생에 따르는 여러 가지 문제점을 일으킬 수 있다. 신용평가데이터는 특정한 시점에서 단순 임의추출하는 것이 아니라 시간이 경과함에 따라 신데이터와 구데이터가 병존하면서 확장해가는 특성이 있기 때문에 시간의 요소까지 고려해야하는 복잡한 구조를 가지고 있다. 따라서 데이터를 전반적으로 관찰하고 이해하는 것은 모형을 구축하는 과정의 가장 기초적이고 중요한 과정이다.

2.1 데이터의 원천

본 연구에서 사용된 데이터는 2005년 2월부터 현재까지 국내의 한 금융회사의 데이터중 일부를 무작위로 추출한 것이며 주요 변수는 CB(Credit bureau)와 관련된 정보로서 국내 3개의 신용평가정보회사에서 종합한 요약정보들로 구성되어있다. 모형 설정을 위한 최소한의 표본수에 대한 명확한 이론적 근거와 제한은 없었으나 안정적이며 표본과 변수들의 다양한 특성이 반영된 평점표를 개발하기 위해서 우량과 불량의 표본수가 충분히 표본내에 존재할 수 있도록 각각 1,500개, 500개 이상이 되도록 추출하였다. 총 3,135개의 관측치중 신용상태가 우량인 관측치는 1,816개, 불량인 관측치는 569개였으며 결측치는 750개가 포함되어 있었다.

데이터의 기간구조는 2005년 2월부터 2008년 1월까지 개발측정기간으로 설정되어 있으며 2008년 1월부터 2009년 1월은 성과 측정기간[2], 2008년 2월 이후부터는 실제 적용 및 검증기간으로 구성되어져 있다. 추출된 자료를 구성하는 변수요약표는 다음과 같다.

2) Basel2 협약(2008년 시행)에 의거하여 측정기간은 1년으로 정의하고 있다(이명식등, 2007).

<표 2-1 > 변수요약표

변수	특성항목	변수	특성항목
YK001	차량가격	YK026	최대 연체금액
YK002	대출신청금액	YK027	최대 연체경험일수
YK003	재직기간	YK028	최초 신용카드 개설일로부터 경과일수
YK004	성별	YK029	최근 신용카드 개설일로부터 경과일수
YK005	연령	YK030	현재 보증건수
YK006	최초 상담일로부터 경과일수	YK031	현재 보증금액의 합
YK007	최근 상담일로부터 경과일수	YK032	최초 보증 발생일로부터의 경과일수
YK008	총자산	YK033	최근 보증 해지일로부터의 경과일수
YK009	순자산(자본총계)	YK034	보증발생건수
YK010	납입자본금	YK035	보증발생금액
YK011	매출액	YK036	신용카드 총 개설건수
YK012	영업이익	YK037	총 신용 개설건수
YK013	분기내 수신신규	YK038	최근 대출일자로부터 경과일수
YK014	은연 채무불이행 해제건수	YK039	총 대출 건수
YK015	은연 채무불이행 해제일로부터의 기간	YK040	총 대출 금액
YK016	총이용금액합계(CA제외)	YK041	최근 조회일로부터 경과일수
YK017	일시불이용금액합계	YK042	최초 조회일로부터 경과일수
YK018	총한도합계금액	YK043	전체 조회건수
YK019	총이용기관수	YK044	전체 조회업체수
YK020	총이용잔액합계	YK045	은행업권 조회건수
YK021	할부이용금액합계	YK046	대부업권 조회건수
YK022	현금서비스이용금액합계	YK047	총 조회건수(은행)
YK023	현금서비스이용기관수	YK048	총 조회건수(상호저축)
YK024	연체 건수	YK049	총 조회건수(카드)
YK025	연체 금액의 합	YK050	총 조회건수(캐피탈)

2.2 변수정의

목표변수는 각 관측치의 신용상태를 기준으로 정의하였으며 우량, 불량, 결측치로 구성되어 있다. 신용상태를 분류하는 기준은 아래의 <표 2-2>와 같다. 상대적으로 불량건수가 부족한 상황에 대비하고 리스크 관리를 강화할 수 있도록 성과측정시점에 연체가 진행중인 관측치에 대해서는 보다 엄격한 기준을 적용하였다. 본 연구에서는 전체 3,135건중 결측치 750건을 제외한 2,385건을 분석에 활용하였다.

<표 2-2> 목표변수 분류기준

구 분	기준	건수	구성비(%)
불 량	· 내부연체 3회이상 · 채무불이행 등록 · 연체기간 60일 이상	569	18.14
우 량	· 내부연체 1회차이하 · 연체기간 1일 ~ 29일 · 연체없음	1,816	57.94
결 측 치 (판단미정)	· 내부연체 2회차 · 연체기간 30일 ~ 59일	750	23.92
계		3,135	100

설명변수들은 앞의 <표 2-1> 변수요약표에 제시된 바와 같이 총 50개로 구성되어 있으며 변수특성별로 개인신상정보로부터 상담관련 정보까지 총 7개의 직군으로 분류할 수 있다.

<표 2-3> 변수특성별 분류기준

특성	변수
개인신상정보	YK001, YK003, YK004, YK005, YK008, YK009
채무불이행	YK014, YK015, YK024, YK025, YK026, YK027
대출관련정보	YK002, YK038, YK039
보증관련정보	YK030, YK031, YK034, YK035
신용카드사용정보	YK016, YK017, YK018, YK019, YK020, YK021, YK022, YK023, YK028, YK029, YK036, YK037
신용조회정보	YK041, YK042, YK043, YK044, YK045, YK046 YK047, YK048, YK049, YK050
상담관련정보	YK006, YK007

위의 <표 2-3> 변수특성별 분류기준에 따라 직관적으로 신용상태에 대한 판단이 가능할 수도 있다. 이를테면 채무불이행 건수가 많을수록, 신용조회 건수가 많을수록, 신용카드 사용정보중 연체, 현금서비스등 불량정보의 빈도수가 높을수록 신용상태가 불량일 가능성이 높을 수 있다. 이러한 개연성을 통계적 검증을 통해 확인할 수 있다면 신용평가모형을 구축하는데 있어서 변수식별 및 판단에 큰 도움이 될 수도 있다.

2.3 독립변수의 영향력

신용평가 모형이 개인 혹은 전문가의 주관과 경험에 의존하고 있다면 모형의 가장 중요한 요소인 객관성과 일관성을 확보할 수 없다. 따라서 본 연구에서는 두 가지 방법으로 독립변수들의 영향력을 살펴보았다. 50개의 설명변수가 모두 포함된 로지스틱 회귀모형의 분석결과와 각 특성별 변수들이 포함된 7개의 로지스틱 회귀모형의 분석결과를 통해 각각의 독립변수들이 신용상태라는 목표변수에 대해 통계적으로 어떠한 영향력을 미치고 있는지 확인해 보았다. 신용상태에 영향을 주는 유의한 주요변수는 아래의 표와 같다.[3] 그러나 <표 2-4>에서 선별된 변수들을 분석결과 그 자체로 받아들일 수는 없다. 분석결과가 통계적으로 타당하게 제시되었다고 하더라도 실제에서 받아들여지기 어려운 경우는 모형이 타당성을 잃기 때문이다. 이를테면 개인신상정보의 변수 중 재직기간의 추정치는 "-0.3378"로 통계적으로 유의하지만 재직기간이 짧을수록, 연령이 적을수록 신용상태가 우량일 가능성이 높다는 것은 상식적으로 받아들여지기 어렵다.[4] 따라서 단순히 로지스틱 회귀분석의 결과를 통해 전체적인 자료를 이해하거나 적절한 모형을 구축하기에는 한계가 있다.

3) 로지스틱회귀분석의 p-value 0.1이하와 오즈비(신뢰구간 95%)를 동시에 유의하게 충족시키는 변수를 선별하였다.

4) 이러한 결과가 발생하는 원인은 변수선택방법, 변수별 상관계수, 다중공선성등 여러 가지 요인을 종합적으로 고려하여 최적의 모형을 선택하는 과정을 거쳐지 못하였기 때문이다. 그러나 이러한 과정들은 변수를 식별하기 위한 기초과정으로서의 연구목적에서 벗어나기 때문에 생략하기로 한다.

<표 2-4> 변수직군별 영향력 변수

변수직군	변수
단순 회귀분석	YK002,YK003,YK005,YK007,YK008,YK010,YK011,YK012, YK013,YK014,YK015,YK018,YK020,YK022,YK023,YK024, YK028,YK030,YK031,YK032,YK034,YK035,YK036,YK037, YK041,YK042,YK043,YK044,YK046,YK049,YK050
개인신상정보	YK003, YK005
채무불이행	YK015, YK024
대출관련정보	없음
보증관련정보	YK030, YK034
신용카드사용정보	YK023, YK037
신용조회정보	YK041, YK043, YK044, YK045, YK046, YK048
상담관련정보	YK007
전체(Full)모형 (50개 변수포함)	YK003, YK004, YK015, YK109, YK023, YK024, YK025 YK039, YK042, YK048

2.4 자료의 정제

분석이 용이한 자료를 구축하기 위한 방안으로 변수선별과 변수변환을 고려할 수 있다. 50개의 설명변수들은 금액, 기간, 조회수등 중복되는 변수들을 제거하거나 변수들간의 상관관계를 분석하여 상관관계가 높은 변수들을 제거하는 과정을 반복하며 최적의 변수를 선별해 낼 수도 있다. 그러나 이러한 과정은 지나치게 복잡하고 시간이 많이 소요되기 때문에 효율성이 떨어지고 원활한 연구진행과 연구목적에도 부합되지 않는다. 변수선별은 모형구축시 SAS E-miner의 Variabel Selection Node를 활용하여 변수를 선별하기로 하고 변수변환위주로 자료를 1차적으로 정제하였다. 각 변수들의 분포를 관찰하며 분포의 형태가 극단적인 변수 위주로 선별하였다. 주로 금액, 경과일수와 관련된 변수들로서 일수는 연단위 기간으로, 금액은 구성비와 변수의 표준편차를 고려하여 순서형으로 변환하였다.

2.5 변수변환의 결과

변환이 완료된 변수를 대상으로 독립변수의 영향력을 2.4절의 방법과 동일하게 로지스틱 회귀분석결과를 적용하여 확인하였다. 변수변환이후 50개의 변수가 모두 포함된 전체(Full) 모형에서는 8개의 변수가 P-value와 오즈비에서 더욱 유의하게 나타났으며 변수직군별 분석결과에서는 43개의 변수중 14개, 단순 로지스틱 회귀분석에서는 19개의 변수가 추가로 유의하게 식별되었다. 따라서 변수변환은 변환 이전에 비하여 신용상태에 영향을 주는 설명변수를 식별하는데 효율적인 방법이 될 수 있으며 자료정제의 목적을 완벽하지는 않지만 어느 정도 달성하였다고 할 수 있다.[5)]

5) <부록 1-1> 변수변환에 따른 변수별 영향력 참조

제 3 장 분석방법론

신용상태를 예측하는 분석방법으로 여러 가지가 있다. 본 장에서는 분석방법론 중 의사결정나무와 회귀분석방법을 활용하여 모형을 구축하는 방법과 과정을 연구하였으며 각각의 분석방법에서 최종적으로 도출된 모형을 비교하였다. 의사결정나무분석은 전문적인 지식이 갖추어져 있지 않더라도 결과를 쉽게 이해할 수 있는 장점이 있는 반면 분리기준이 되는 근방에서 예측오류가 클 가능성이 있으며 선형효과가 결여된다는 한계점이 있다. 회귀분석은 가장 널리 통용되고 있는 분석방법으로 해석이 편리하고 유용한 정보를 많이 얻을 수 있지만 선형성을 가정하고 있기 때문에 비선형성을 가지는 경우에는 많은 한계점을 드러내기도 한다 (강현철등, 2007).

3.1 의사결정나무분석

의사결정나무(Decision Tree)의 가장 큰 장점은 회귀분석(Regression Analysis), 신경망(Neural Networks)에 비해 분석과정을 쉽게 이해하고 설명할 수 있다는 것이다. 의사결정나무는 글자 그대로 의사결정규칙을 나무의 마디(node), 가지(branch), 잎(leaf)의 구조를 토대로 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석방법이다. 의사결정나무의 또 다른 장점은 회귀분석과 같은 모수적 모형 분석의 사전단계로써 이상치(outlier)를 검색하거나 분석에 필요한 변수들을 잠정적으로 판단하는데 큰 도움이 된다는 것이다.

본 연구에서는 지도예측방법(Unsupervised modeling)을 바탕으로 예측, 차원축소 및 변수의 선택에 중점을 두었다. 세분화, 교호작용의 파악, 범주의 병합 및 연속형 변수의 이산화등과 같은 자율예측방법(Supervised modeling)을 활용하여 연구를 진행하며 비교분석할 수도 있었지만 데이터 변환 및 정제 과정에서 선별되었기 때문에 제외하였다.

3.1.1 분석절차 및 흐름도

의사결정나무는 분석의 목적과 자료의 구조에 따라서 적절한 분리기준과 정지규칙을 지정하면서 형성되어간다. 분류오류를 크게 할 수 있는 위험이 높거나 부적절한 추론규칙을 가지고 있는 경우에는 가지치기를 통하여 가지를 제거하고 이익도표나 위험도표등과 같은 모형평가 도구, 검증용 자료에 의한 교차타당성등을 이용하여 의사결정나무를 평가한다. 본 연구에서는 자료의 정제과정을 통해서 변환된 변수들로 구성된 데이터마트를 활용하였다.

변수변환 못지 않게 중요한 과정이 결측치를 제거하는 것이다. 변수에 결측치가 많다면 결측치 자체를 하나의 관측값으로 이해하여 왜곡된 결과를 초래할 수 있기 때문에 결측값에 대한 적절한 조치는 매우 중요하다. 본 연구에서는 Variable Selection Node를 통하여 1차적으로 변수를 선별하였고 Tree Node의 Variables Tab에서 중복되고 불필요하다고 판단되는 변수를 제거하고 모형을 평가하였다. 모형 평가에서 사용할 변수를 선별하는 과정은 체계적이고 논리적인 절차에 의해서라기보다는 경험적인 요소를 토대로 평가결과를 수차례 분석하는 시행착오를 반복하면서 최종 변수를 식별하였다. 데이터의 분할은 Train : Validation : Test의 비율을 4 : 3 : 3으로 지정하였고 분할방법은 단순임의추출로 지정하였다. 분리기준은 SAS E-miner에 지정된 3가지 방법을 모두 사용하였다.

3.1.2 분석알고리즘과 분리기준

의사결정나무분석에서 사용되는 분리기준은 Chi-square Test[6], Entrophy Index[7], Gini Index[8] 3가지의 방법을 사용한다. Chi-square Test는 P-value가 가장 작은 입력변수와 그때의 최적 분리에 의해서 자식마디를 형성한다. Entropy Index는 다항분포에서의 우도비 검정통계량을 사용하는 것으로 Entropy Index가 가장 작은 입력변수와 그때의 최적분리에 의해서 자식마디를 형성하며 Gini Index는 Gini Index 값을 가장 크게 감소시켜주는 입력변수와 그때의 최적분리에 의해서 자식마디를 선택한다. Chi-square 통계량이 Gini Index나 Entropy

---

6) $X^2 = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$: CHAID, Kass (1980)

7) $G = -\sum_{j=1}^{c} \sum P(i)P(j)$: C4.5, Quinlan (1993)

8) $\sum\sum P(i)P(j) = \sum P(j)(1-P(j)) = 1 - \sum P(j)^2 = 1 - \sum (n_j/n_o)^2$: CART, BFSO (1984)

Index에 비해서 보다 단순한 형태의 나무구조를 가지게 하는 경향이 있다 (최종후 등, 2001).

3.1.3 분리기준별 모형결과

의사결정나무분석 모형을 3가지의 알고리즘과 분리기준에 따라 의사결정나무의 구조와 특징을 확인하였다. 객관적인 비교를 위해 각 모형에서 동일한 데이터마트와 입력변수를 사용하였으며 전체 50개의 변수중 의사결정나무 모형을 구현하는데 필요하다고 판단되는 변수들을 선별하기 위해 입력오류, 범위초과, 필드간 중복 및 상충되는 성질의 변수들을 파악하여 제거하였다. 50개의 변수 중 20개의 변수가 잡음을 발생시키는 데이터로 발견되어 모형구축에는 30개의 변수로 구성된 데이터마트를 구축하였다. 모형의 평가는 각각의 방법에 대해 사전확률과 모형 구축후 분류된 사후확률을 비교하여 모형의 분류 효용성과 타당성을 평가하였고 정오분류표를 토대로 모형이 목표변수를 얼마나 정확하게 분류하고 있는지 확인하였다. 각 모형의 예측성능 평가를 위해 알고리즘에 따른 ROC curve와 C-통계량을 비교하였다.

① 정오분류표

Chi-square : CHAID 알고리즘

이익행렬도표를 통해 각각의 분류 기준값에 따른 정오분류표, 사전분포와 사후분포의 정분류율은 아래의 <표 3-1>에 제시된 바와 같다. 전체적인 측면에서는 사전분포의 정분류율과 사후분포의 정분류율을 비교하여 확인할 수 있다. "Thresshold=50%"로 설정하였을 경우 실제 신용상태가 우량일 때 우량으로 예측할 확률이 95.89%를 오류없이 분류해낼 수 있다. 반면에 불량인 경우는 전체적으로 7.5%, 불량 범주에 대해서는 30%를 오류없이 분류해 낼 수 있다. 따라서 X^2 통계량을 사용한 모델은 신용상태가 우량 범주일 때 우량으로 분류하는 모델로 활용할 수 있다. 반대로 신용상태가 불량 범주일 때 불량으로 분류하는 최적 모델로는 적합하지 않지만 사전분포에 비하여 정분류할 수 있는 확률이 4배이상 높아졌음을 확인할 수 있다. 그러나 불량을 우량으로 예측하는 오분류율은 치명적일 수 있기 때문에 정분류율이 높고 우량을 불량으로 오분류하는 확률이 작다고 할지라도 반드시 확인할 필요가 있다. X^2모형의 경우 불량을 우량으로

예측하는 오분류율은 17.5%로 비교적 높다고 할 수 있다. 따라서 모형의 전체적인 정확성과 정분류율을 높이는 방법도 중요하지만 치명적인 오분류을 줄이면서 전체적인 안정성을 높이는 것이 더욱 중요하다. 안정성은 이익행렬도표의 분류기준값을 조정하면서 최적의 값을 확인할 수 있다. X^2통계량을 활용한 모형의 경우 Thresshold를 65% ~ 85% 사이로 설정할 때 안정성이 비교적 높아지는 것을 확인할 수 있다. X^2를 사용한 의사결정나무 모형의 결과는 다음과 같다.

이익행렬도표는 기준값을 50%로 설정하였을 때의 오분류율과 정분류율을 포함하고 있다. 정분류율은 79.3%로 사전분포에 비해 정분류율이 5.3% 향상되었으나 정오분류율이 향상이 안정성을 반드시 보장하는 것은 아니기 때문에 X^2통계량을 사용한 모형이 안정성에 대해서는 추가적인 분석이 요구된다.

<표 3-1> CHAID 알고리즘의 정오분류표

예측 / 실제	Thresshold=50			Thresshold=65		
	1 (우량)	0 (불량)	합계	1 (우량)	0 (불량)	합계
1 (우량)	514 (80%)	22 (29%)	536	424 (90%)	112 (46%)	536
0 (불량)	126 (20%)	54 (71%)	180	50 (10%)	130 (54%)	180
합계	640	76	716	474	242	716
정분류율	· Prior Distribution : 74.8% · Post Distribution : 79.3%			· Prior Distribution : 74.8% · Post Distribution : 77.3%		
오분류율	· Prior Distribution : 25.2% · Post Distribution : 20.7%			· Prior Distribution : 25.2% · Post Distribution : 22.7%		

Entropy Index : C 4.5 알고리즘

Entropy Index를 사용한 의사결정나무 모형의 결과는 X^2통계량을 사용한 모형의 결과와 거의 유사하다. 정분류율은 Train의 경우 77.4%에서 83.4%로 6% 정분류율이 높아졌으며 Valid의 경우 74.8%에서 80.5%로 5.7% 정분류율이 향상되었다. X^2통계량을 사용한 모형보다는 전체적으로 정분류율이 향상되었음을

알 수 있다. Entropy Index를 활용한 모형은 우량의 경우 97.7%를 우량으로 분류할 수 있으며 불량의 경우 29.4%를 불량으로 정확하게 분류해낼 수 있다. Entropy Index를 활용한 모형에서 불량을 우량으로 예측할 확률이 동일한 분류값 기준에서 CHAID 알고리즘을 활용한 앞의 모형에 비해 10.4% 낮아졌음을 알 수 있다.

<표 3-2> C 4.5 알고리즘의 정오분류표

예측 / 실제	Thresshold=50			Thresshold=60		
	1 (우량)	0 (불량)	합계	1 (우량)	0 (불량)	합계
1 (우량)	524 (80%)	12 (18%)	536	466 (86%)	70 (40%)	536
0 (불량)	127 (20%)	53 (82%)	180	76 (14%)	104 (60%)	180
합계	651	65	716	542	174	716
정분류율	· Prior Distribution : 74.8% · Post Distribution : 80.5%			· Prior Distribution : 74.8% · Post Distribution : 79.6%		
오분류율	· Prior Distribution : 25.2% · Post Distribution : 19.5%			· Prior Distribution : 25.2% · Post Distribution : 20.4%		

Gini Index : CART 알고리즘

Gini Index를 사용한 의사결정나무 모형 역시 X^2나 Entropy Index를 사용한 모형과 결론은 흡사하다. 그러나 Gini Index를 사용한 의사결정나무 모형은 앞서 제시한 두가지 방법들에 비하여 우량범주를 제대로 분류하는 정분류율은 낮으나 불량범주의 변수를 불량으로 예측하는데 있어서는 오분류율이 두 모형에 비해서 훨씬 낮게 나타났다.

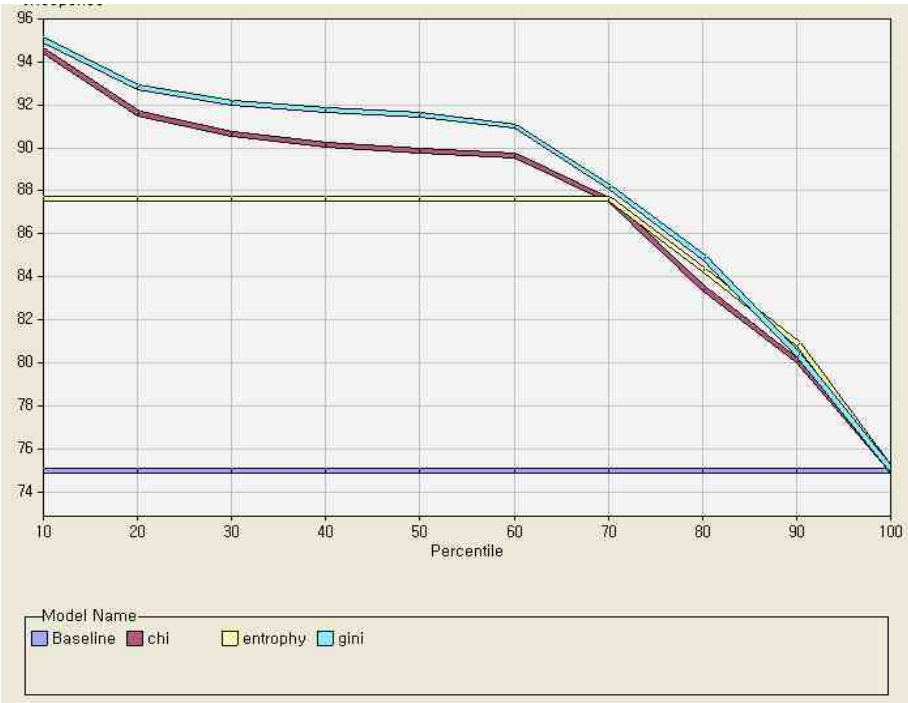
<표 3-3> CART 알고리즘의 정오분류표

예측 / 실제	Thresshold=50			Thresshold=70		
	1 (우량)	0 (불량)	합계	1 (우량)	0 (불량)	합계
1 (우량)	503 (84%)	33 (29%)	536	440 (88%)	96 (44%)	536
0 (불량)	98 (16%)	82 (71%)	180	59 (12%)	121 (56%)	180
합계	601	115	716	499	217	716
정분류율	· Prior Distribution : 74.8% · Post Distribution : 81.7%			· Prior Distribution : 74.8% · Post Distribution : 78.4%		
오분류율	· Prior Distribution : 25.2% · Post Distribution : 18.3%			· Prior Distribution : 25.2% · Post Distribution : 21.6%		

세가지 알고리즘에 따른 정오분류표의 결과는 각 분류기준값에 따라 정오분류율이 다르기 때문에 위에 제시된 정오분류표의 결과를 토대로 객관적인 비교를 하기에는 제한이 있다. 따라서 어느 한 방법이 절대적으로 우위에 있다고 평가하기 위해서는 다른 평가방법의 결과들을 종합적으로 고려할 필요가 있다.

② Lift Graph와 ROC curve

아래의 <그림 3-1> 에서 볼 수 있듯이 세가지 모형 모두 기준선보다 상당히 멀리 떨어져 있다. Lift graph를 기준으로 예측성능을 판단한다면 “Gini > Chi-square > Entropy" 순이 된다.

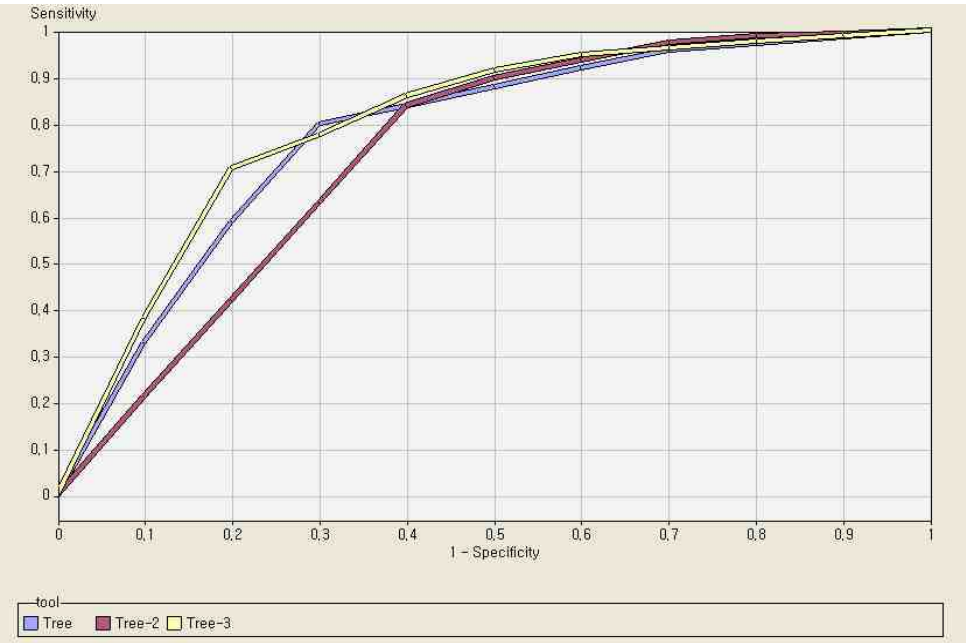


<그림 3-1> 의사결정나무모형의 Lift Chart

ROC curve로 모형을 평가 및 비교하기 위해서는 곡선 하부의 면적을 계산하여야 한다. 계산된 면적은 C 통계량으로 표현하며 0.5~1사이의 값을 갖는다. C-통계량 값이 1에 가까울수록 모형은 좋은 예측성능을 가지고 있다고 할 수 있다. 아래에 제시된 <표 3-4>의 C-통계량을 기준으로 모형의 예측력을 평가한다면 “Gini > Chi square > Entropy" 순이된다[9].

<표 3-4 > 알고리즘별 C-통계량

구분	Chi Square	Entropy	Gini
C-통계량	0.8633	0.8296	0.8830



<그림 3-2 > 의사결정나무의 ROC curve

③ 절단값에 따른 비교

각 방법별 정확도, 민감도, 특이도를 종합한 결과는 아래의 <표 3-5>와 같다. 정확도 측면에서는 Entropy 방법이 전체적으로 큰 변화없이 일관성있는 결과를 제시하고 있다. 그러나 불량을 불량으로 예측하는 특이도에서 다른 모형에 비해 예측성능이 떨어진다. 따라서 Entrophy 모형은 신용상태가 불량인 관측치를 불량으로 예측하는 성능이 다른 두가지 방법에 비해 낮다. 정확도, 민감도, 특이도의 전체적인 관점에서 본다면 “Chi-square > Entropy > Gini" 순이 된다.

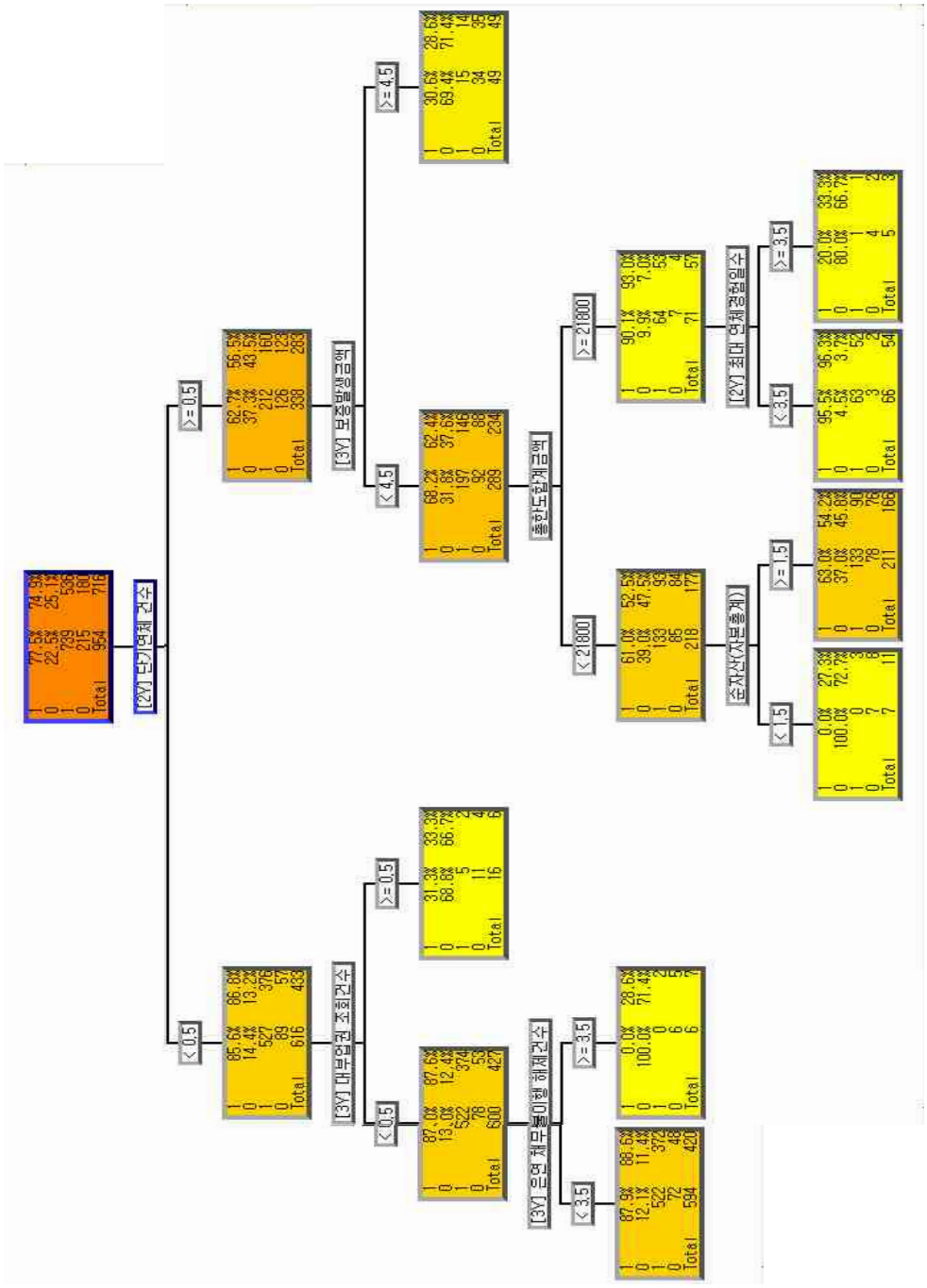
9) 곡선 아래의 면적을 계산한 값을 C-통계량이라고 하며 0.5 ~ 1.0사이에 존재하며 1.0에 가까울수록 예측모형의 성능이 좋다. (강현철등, 2007)

<표 3-5 > 알고리즘별 정확도, 민감도, 특이도

절단값	Chi square			Entrophy			Gini		
	정확도	민감도	특이도	정확도	민감도	특이도	정확도	민감도	특이도
65		0.79	0.72		0.85	0.59	81.44	0.92	0.49
70		0.79	0.72		0.85	0.59		0.82	0.67
75	77.37	0.79	0.72	78.63	0.85	0.59	78.35	0.82	0.67
80		0.79	0.72		0.85	0.59		0.82	0.67
85		0.79	0.72		0.85	0.59	74.58	0.76	0.71

3.1.4 최종모형의 선택과 해석

신용상태를 예측하는 모형에서 모형의 오분류가 불가피한 것이라면 오분류율을 최소화하는 모형이 가장 바람직한 모형이다. 신용상태의 오분류는 우량을 불량으로 예측하는 것보다 불량을 우량으로 예측하는 경우가 더 위험하기 때문에 불량을 우량으로 예측하는 오분류율에 더욱 중점을 둘 필요가 있다. 오분류율이 지나치게 높다면 그 모형은 분류모형으로서 가치를 잃게된다. 지금까지 제시한 여러 가지 평가방법중 예측성능이 가장 높은 모형은 Gini Index를 활용한 모형이다. Gini Index를 통해 구현된 의사결정나무 모형은 그림<3-3>과 같다. 의사결정나무 모형을 해석해보면 단기연체건수가 없고 대부업계 신용정보조회 경험이 없으며 은행연합회 채무불이행 해제건수가 3건 이하일 경우 신용상태가 우량일 가능성이 높게 나타나고 연체건수가 1건 이상이고 보증발생금액이 4.5건 이상인 관측치는 신용상태가 좋지 않을 가능성이 높다. 또한 신용카드 총한도 합계금액과 순자산, 최대연체경험일수도 신용상태의 좋고 나쁨을 판단할 수 있는 기준이 된다.



<그림 3-3> Gini Index 알고리즘을 활용한 최종모형

3.2 로지스틱 회귀분석

회귀분석의 가장 큰 장점은 목표변수가 여러 가지 입력변수들에 의해서 어떻게 설명 또는 예측되는지 함수의 형태로 표현가능하다는 것이다. 회귀분석은 최근의 대다수 평점 시스템을 대표하는 방법중 하나이다. 회귀분석이 선호되는 이유는 다른 분석방법들에 비해 안정적이므로 그 결과로 발생한 모형의 평점에 포함되는 변수의 종류와 변수들의 상대적 가중치 및 상호관계에 대해 완전하게 이해할 수 있으면서도 비교적 쉬운편에 속하기 때문이다. 회귀분석은 목표변수와 일련의 입력변수들을 필요로하게 되는데 이때 입력변수들은 다양한 형태를 지닐 수 있다. 가장 보편적이고 원칙적인 방법은 연속형 변수에 대해서는 원형 데이터를 사용하고 범주형 변수에 대해서는 더미변수를 만드는 것이다. 그러나 이러한 방법은 효율적인 모형을 만드는데 큰 방해요인이 될 수도 있다. 따라서 데이터의 전체적인 분포를 분석하고 적절한 변수변환을 통하여 시의적절한 모형을 구현해내야한다. 의사결정나무분석과 객관적인 비교를 위해서는 동일한 형태의 변수를 사용해야하므로 데이터 정제 단계에서 변환된 데이터를 사용하였으며 목표변수가 '신용상태의 우량, 불량'이라는 이항분포 형태로 나타나기 때문에 로지스틱 회귀분석방법을 사용하였다. 로지스틱 회귀분석 방법을 활용하여 궁극적으로 밝히고자 하는 것은 의사결정나무분석을 토대로 확인된 변수들의 정보를 바탕으로 목표변수에 대한 입력변수들의 수학적 표현을 구체화시키고 의사결정나무분석에서 밝히지 못한 입력변수들의 목표변수에 대한 영향력을 살펴보는데 있다.

3.2.1 분석절차 및 흐름도

의사결정나무분석방법과 동일하게 데이터마트를 확보하고 4 : 3 : 3의 비율로 Train, Validation, Test의 데이터세트를 분할하였다. 준비과정과 분석과정, 최종 선택된 모형의 해석과 비교를 통하여 신용상태에 영향을 주는 입력변수들이 어떠한 것들이 있는지 확인하였다. 변수선택의 기준은 어느 한 기준에 절대적으로 의존하기 보다는 여러가지 방법을 종합적으로 고려하였다. 결측치 발생에 따른

모형의 왜곡을 방지하기 위해 Replacement node를 활용하여 변수별로 보정(imputation)하여 대체값을 산출 및 적용하였고 불필요한 노력의 방지를 위해 Variable selection node를 추가하여 50개의 변수 중 12개의 변수를 제거하여 총 38개의 입력변수를 선별하였다. 선별된 변수들은 회귀분석을 통해 각 변수선택별 모형을 평가하여 가장 적합한 모형을 찾아내고자 하였다.

3.2.2 변수선택방법

50개의 변수를 모두 포함하는 회귀모형을 구현해내는 것은 매우 어려운 일이다. 모든 변수들이 변수선택 기준을 충족시켜 모형에 포함될 수도 있지만 계산하는 것이 매우 복잡할 뿐만 아니라 모형의 수도 기하급수적으로 늘어나게 된다. 따라서 각각의 모형에 대해 검토와 확인을 한다는 것은 거의 불가능하다. 아무리 정확한 모형이라고 하더라도 지나치게 복잡하여 해석이 어렵다면 그 모형은 실질적인 가치를 잃는다고 할 수 있다. 따라서 적절한 방법을 통하여 변수를 선택하는 것이 훨씬 합리적이라고 할 수 있다(김민정등, 2001).

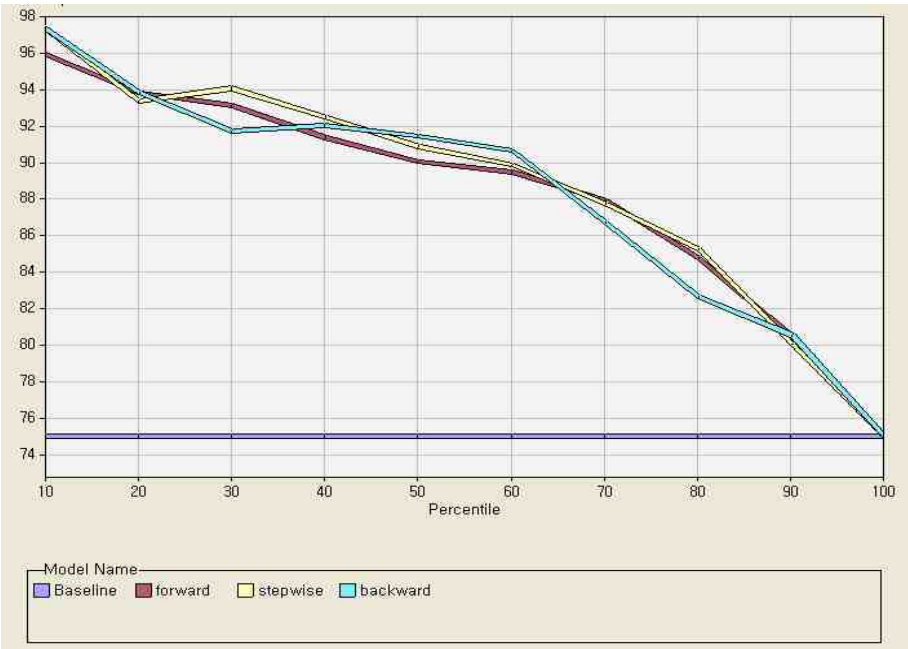
본 연구에서는 다음에 제시된 3가지 방법을 선택하고 비교하여 유효한 변수를 추적하였다. 전진선택방법은 계산이 빠르다는 장점이 있지만 한 번 선택된 변수는 절대로 제거되지 않는다는 단점이 있다. 이러한 경우 각 변수들간에 다중공선성이 존재하거나 경험상 불필요한 변수로 인식된다고 하더라도 모형에서 제거될 수 없는 한계가 있다. 이를테면 성별이라는 변수가 여러변수와 상호작용으로 인하여 가장 강력한 변수로 선택되었다고 한다면 신용상태라는 목표변수는 다른 변수들과 차후에 선택될 변수들에 상관없이 목표변수를 좌우하는 영향력이 강한 변수로 작용하게 되어 분석 결과를 왜곡시킬 우려가 있다. 변수의 수가 많을수록 이러한 왜곡 가능성은 더욱 높아질 수 있다. 후방제거법은 중요한 변수가 모형에서 제외될 가능성이 적으므로 비교적 안정적인 방법이지만 한 번 제외된 변수는 다시 선택되지 못한다는 한계가 있다. 단계적방법은 전방선택법과 후방제거법을 적절히 조율한 효율적인 방법이다.

3.2.3 변수선택방법별 모형결과

① Lift graph

Lift graph는 예측된 사후확률을 이용하여 모형을 평가하는 방법이다. 아래에 제시된 <그림 3-4> Lift graph의 횡축은 데이터세트의 개체들을 등급화한 수치이고 종축에 표시된 반응률은 각 등급에서 목표변수인 신용상태 우량의 비율을 의미한다. 각 등급의 Lift를 살펴보면 기준선 반응률에 비해 각 등급에서의 반응율이 높게 나타나고 있으며 세가지 방법 모두 상위 3등급까지는 기준보다 1.25배 이상 반응률이 높게 나타나고 있다. 전체적으로 회귀모형은 각 등급에 따른 반응률의 차이를 보이고 있으며 상위등급에서 하위등급으로 갈수록 반응률이 낮아지고 있으므로 예측모형의 비교적 좋다는 것을 알 수 있다.

기준선으로부터 멀리 떨어져 있을수록 모형의 예측성능이 좋다고 할 수 있는데 아래의 <그림 3-4>에서는 “단계적방법 > 전방선택법 > 후방제거법”순이다.



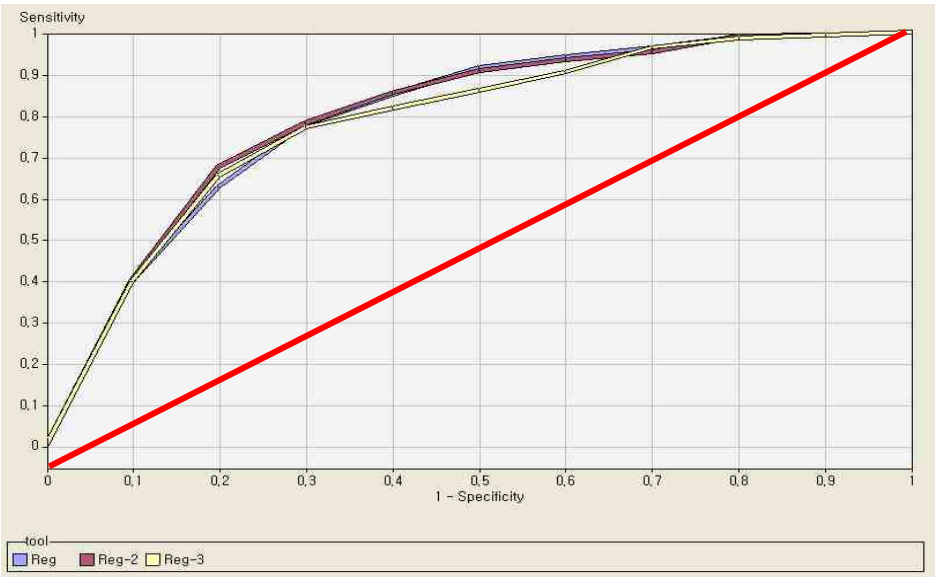
<그림 3-4> 로지스틱 회귀분석모형의 Lift Graph

② ROC Curve

ROC Curve는 예측모형의 성능을 도표로 평가하기 위한 방법으로 정확도 측면에서 모형의 성능을 평가할 수 있다. 아래의 <그림3-5>에서 대각선은 우연에 의한 ROC을 의미한다. 모형의 성능을 비교하는 기준은 ROC의 면적이다. 대각선이 면적은 전체면적의 0.5(50%)이다. 따라서 곡선이 대각선보다 상부에 위치하고 있고 곡선이 차지하는 면적이 0.7보다 크므로 예측모형의 성능이 좋다고 판단할 수 있다. 모형별로 예측성능을 판단해보면 단계적방법과 후방제거법이 전방선택법에 비하여 성능이 좋고 단계적 방법과 후방제거법의 성능우열은 그래프를 통해 확인하기에는 제한되지만 C-통계량 값을 비교하여 확인할 수 있다. 아래의 <표 3-6>에서 “단계적방법 > 전진선택법 > 후방제거법” 순으로 예측성능이 좋다고 할 수 있다.

<표 3-6 > 변수선택방법별 C-통계량

구분	전진선택법	후방제거법	단계적방법
C-통계량	0.8738	0.8637	0.8896



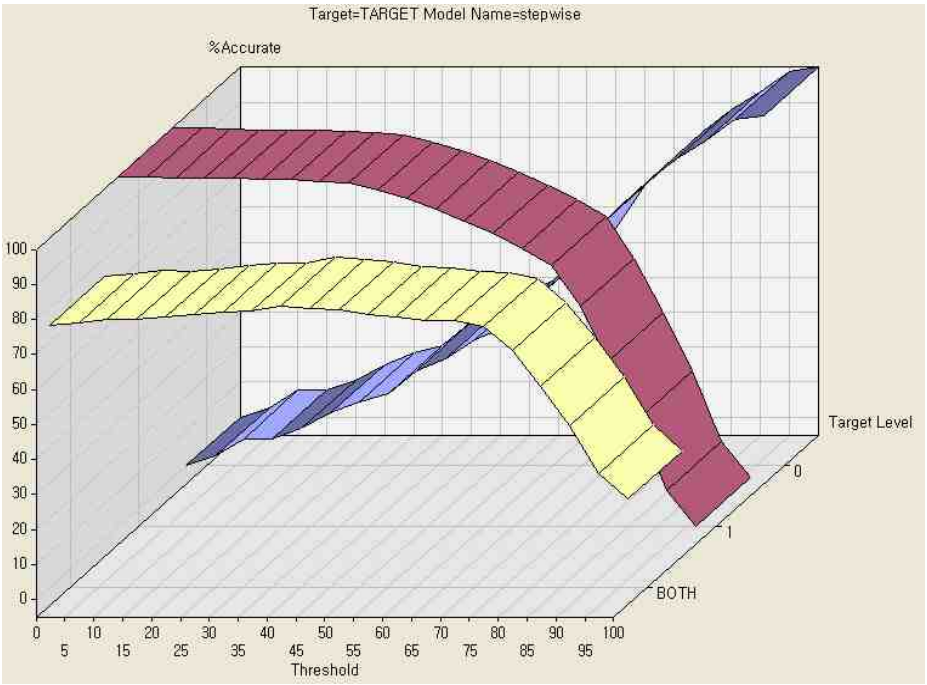
<그림 3-5 > 로지스틱 회귀분석모형의 ROC Curve

③ 절단값(Cut-off value)에 따른 비교

각 변수선택별 정분류율이 가장 높은 방법은 절단값에 따라 각각 다르다. 각 방법별로 정확도에 초점에 지나치게 초점을 맞추면 분석결과를 왜곡시킬 수 있다. 이를테면 신용상태가 우량인 데이터의 수가 불량인 데이터의 수보다 몇 배이상 많은 경우 신용상태 불량에 대한 오분류율은 은폐되기 쉽다. 따라서 민감도와 특이도를 적절히 확인하며 절단값을 선택해야 한다. 아래의 <표 3-7>을 보면 정확도가 가장 높은 경우 민감도 역시 높지만 특이도가 매우 낮다. 신용상태가 불량인 데이터를 불량으로 바르게 예측하지 못하고 우량으로 잘못 예측한다는 것으로 위험 관리 측면에서는 부적절하다. 적어도 신용상태가 불량인 데이터를 불량으로 예측할 수 있는 정도가 50%는 넘어야 최소한의 위험관리가 가능하고 할 수 있다. 각 방법별로 본다면 전진선택법의 경우 절단값은 60~75%가 적당하고 후방제거법과 단계적방법의 경우는 65~75%가 적절하다.

<표 3-7 > 절단값에 따른 변수선택방법별 정확도, 민감도, 특이도

절단값	전진선택법			후방제거법			단계적방법		
	정확도	민감도	특이도	정확도	민감도	특이도	정확도	민감도	특이도
40	0.795	0.974	0.261	0.803	0.981	0.272	0.803	0.994	0.267
45	0.796	0.966	0.289	0.796	0.961	0.306	0.796	0.986	0.289
50	0.797	0.944	0.361	0.792	0.937	0.361	0.777	0.964	0.350
55	0.811	0.938	0.433	0.778	0.907	0.394	0.773	0.904	0.417
60	0.810	0.903	0.533	0.774	0.873	0.478	0.762	0.862	0.428
65	0.791	0.860	0.583	0.763	0.838	0.539	0.761	0.816	0.589
70	0.777	0.813	0.667	0.761	0.795	0.661	0.744	0.770	0.672
75	0.743	0.746	0.733	0.744	0.746	0.739	0.655	0.653	0.744



<그림 3-6 > 단계적방법 Correct Classfication

④ 주요통계량

각 방법에서 산출된 통계량을 통해서도 모형 성능을 판단할 수 있다. <표 3-8>은 변수선택방법별로 산출된 주요 통계량이다. 아래에 제시된 통계량들은 모두 값이 작을수록 모형의 성능이 좋다고 할 수 있다. 따라서 4개의 주요통계량을 종합해 볼 때 모형의 성능은 “단계적방법 > 전진선택법 > 후방제거법”순이 된다.

<표 3-8 > 변수선택방법별 주요통계량

구분	전진선택법	후방제거법	단계적방법
AIC	840.9	854.6	841.3
ASE	0.1396	0.1442	0.1393
SBC	889.5	898.4	885.1
MISC	0.2025	0.2081	0.2025

3.2.4 최종모형의 선택과 해석

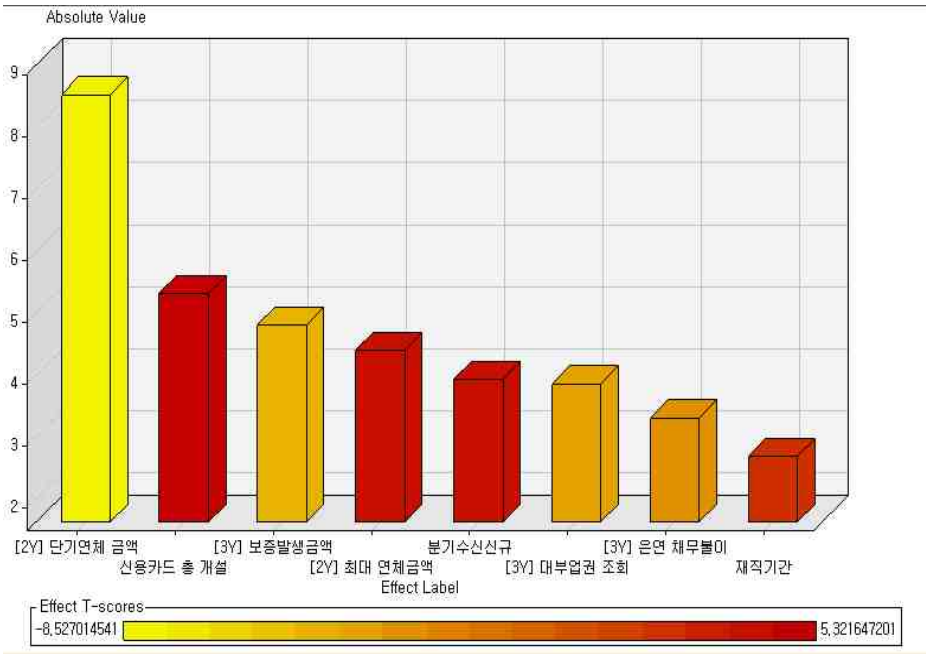
단계적방법이 다른 두 방법에 비해 절대적으로 예측성능이 우수한 것은 아니지만 지금까지 결과를 종합하였을때 가장 예측성능이 좋다고 할 수 있다. 단계적방법을 바탕으로 선택된 변수와 각 변수별 영향력은 <표 3-9>에 제시되어있다. 재직기간, 분기수신 평잔, 신용카드 개설건수가 많을수록 신용상태가 좋을 가능성이 있으며 은행연합회 채무불이행 해제건수가 적을수록 현금서비스 이용기관수, 단기연체금액, 보증금액과 대부업권 조회수가 작을수록 신용상태가 우량일 가능성이 높다.

모수를 바탕으로 변수들의 영향력을 판단하기 위해서 effect-score를 산출하였는데 결과는 <그림 3-7>과 같다. 변수별로 영향력을 살펴보면 단기연체금액의 합이 신용상태에 가장 큰 영향을 주는 것으로 분석되었고 신용카드개설건수, 보증발생금액, 최대연체금액, 분기내수신신규, 대부업권조회수, 은행연합회 채무불이행건수, 재직기간 순으로 신용상태에 영향력을 주고 있다고 할 수 있다. 단계적방법에 의해 식별된 9개의 변수를 활용하여 구현된 최종 모형의 수식은 다음과 같다.

$$Logit(\pi) = 0.0455x_1 + 0.5898x_2 - 0.3323x_3 - 0.4654x_4 \\ - 0.5547x_5 - 0.4560x_6 - 0.3326x_7 + 0.2303x_8 - 1.9674x_9$$

<표 3-9> 단계적방법 모수추정치와 T-scores

Effect Name		Effect Label	Parameter Estimate	Effect T-scores	Effect Sign
X1	YK003	재직기간	0.0455	2.6863	+
X2	YK013	분기수신평잔	0.5898	3.9380	+
X3	YK014	은연 채무불이행 해제건수	-0.3323	-3.3167	-
X4	YK023	현금서비스이용기관수	-0.46549	-2.5571	-
X5	YK025	단기연체 금액의 합	-0.5547	-8.5270	-
X6	YK026	최대 연체금액	-0.4560	4.3989	-
X7	YK035	보증발생금액	-0.3326	-4.8118	-
X8	YK036	신용카드 총 개설건수	0.2303	5.3216	+
X9	YK046	대부업권 조회건수	-1.9674	-3.8665	-



<그림 3-7> 단계적방법 모형 effect-score

3.3 두 방법론의 비교

본 절에서는 의사결정나무분석방법과 로지스틱 회귀분석방법을 비교하고 각 방법론에 대해 최적화된 모형을 비교하여 어떤 모형의 성능이 우수한지 살펴보았다. 의사결정나무분석방법에서 예측성능이 가장 우수한 CART 알고리즘을 사용한 모형과 로지스틱 회귀분석방법에서 예측성능이 가장 우수한 단계적 회귀분석 방법을 선택하여 비교하였다. 예측성능을 평가하는 기준으로는 Lift Graph, ROC Curve 두가지 방법을 활용하였다.

3.3.1 Lift Graph

첫 번째로 두 모형의 Lift Graph를 비교해보았다. 그래프 하단의 75%는 모형을 적용하기 전 무작위로 관측치들을 관찰하였을때의 반응율을 의미한다. 로지스틱

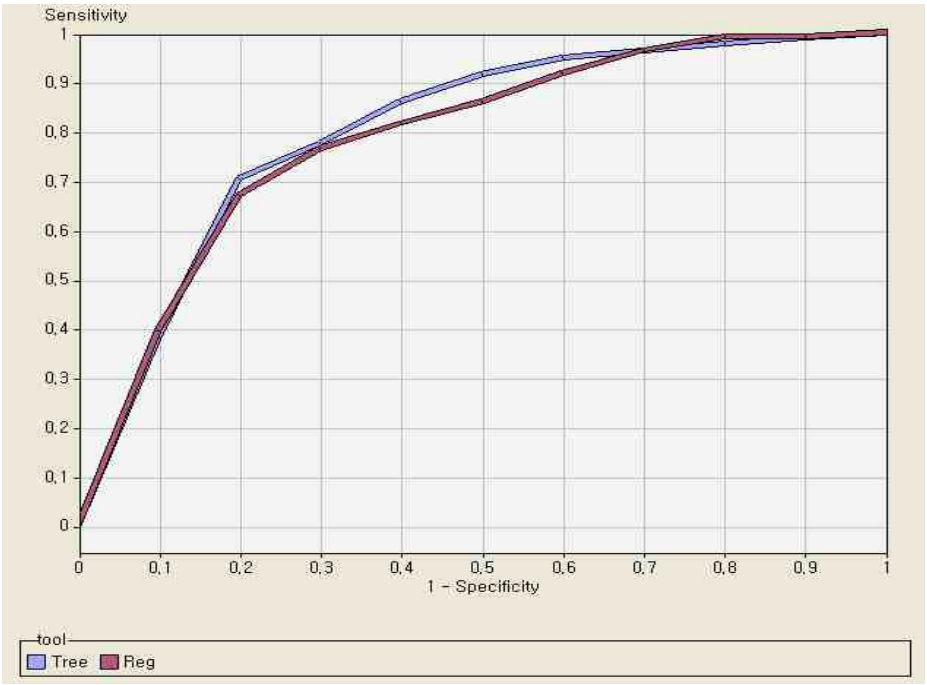
회귀분석방법의 경우 상위 10% 관측치에서의 반응율은 97%, 의사결정나무분석 방법에서는 95%의 반응율을 보여주고 있다. 상위 10% ~ 30% 범위에서는 로지스틱 회귀분석이 의사결정나무분석에 비하여 기준선으로부터 멀리 떨어져 있고 30% ~ 60%까지는 두 방법이 비슷한 수준을 나타내고 있으며 60% ~ 90%의 범위에서는 의사결정나무분석방법이 로지스틱 회귀분석방법에 비하여 멀리 떨어져 있음을 확인할 수 있다.

<그림 3-8> 의사결정나무분석과 로지스틱 회귀분석의 Lift Graph

3.3.2 ROC Curve

Lift Graph에 비해 ROC Curve는 비교적 예측성능을 육안으로 식별하기에 용이하다. 일반적으로 좋은 성능을 가진 모형일수록 ROC Curve는 대각선 위쪽에 위치하게 된다. 아래의 ROC Curve에서 로지스틱 회귀분석의 Curve가 의사결정나무분석의 Curve보다 하부면적이 넓다. 실제로 C-통계량을 산출해보면 로지스틱 회귀분석의 C-통계량이 의사결정나무분석의 C-통계량보다 0.006정도

높았다. 따라서 ROC Curve를 통해서는 의사결정나무분석방법보다 로지스틱 회귀분석 모형이 비교적 성능이 좋다고 할 수 있다.



<그림 3-9> 의사결정나무분석방법과 로지스틱 회귀분석 ROC

두 모형을 비교해본 결과 모형의 예측성능 차이는 미미하지만 로지스틱 회귀분석방법의 성능이 우수하였다. 뿐만 아니라 모형의 안정성, 해석의 용이성과 실무에서의 활용 가능성등을 고려해볼 때 로지스틱 회귀분석방법을 바탕으로 신용평점 모형을 산출하는 것이 바람직하다.

제 4 장 신용평점표 산출

지금까지 분석한 결과를 토대로 신용평점표를 작성하였다. 일반적으로 신용평점표를 산출하는 절차는 자료를 추출하고 정리하는 준비절차, 초기 특성변수를 분석하여 신용평점표를 산출하는 추정절차와 신용평점표의 타당성과 적합성을 검증하는 절차로 구분될 수 있다.

제 2장과 제 3장에서 연구한 자료의 정제와 분석방법론은 신용평점표를 작성하기 위한 준비절차라고 할 수 있다. 본 장에서는 분석방법론의 결과를 바탕으로 최종 신용평점표를 작성하고 평점표의 타당성을 평가하였다.

4.1 신용평점표 산출절차

4.1.1 적용모형

의사결정나무분석과 로지스틱 회귀분석의 결과를 비교하여 상대적으로 예측성능이 우수한 로지스틱 회귀분석방법을 신용평점표 산출 방법으로 결정하였다. 로지스틱 회귀분석을 선택한 이유는 상대적으로 우수한 예측성능 이외에도 실무에서 가장 많이 적용되고 있을 뿐 아니라 자료정제과정을 통해 이산화, 범주화된 변수들을 활용할 수 있는 적합한 방법이기 때문이다. 의사결정나무분석의 경우 각 마디별 가중치화된 수치를 부여하여 평점표를 산출하는 방법을 고려해 볼 수도 있겠으나 하부마디가 상부마디에 종속됨에 따라 상관계수가 높고 다중공선성이 존재할 가능성이 높다. 또한 변수가 이진형으로 단순하게 분할되어 변수별 특성과 가중치를 적절하게 반영할 수 없다는 한계점도 있다.

4.1.2 변수선택

신용평점표를 작성하는데 포함되는 변수는 로지스틱 회귀분석의 3가지 변수선택의 결과를 종합적으로 고려하였다. 전진선택방법, 후방제거법, 단계적방법의 모형에서 유의한 변수들을 모두 포함하는 새로운 로지스틱 회귀모형의 결과를 바탕으로 평점표산출에 포함될 최종적인 변수들을 선별하였다[10]. 신용평점표 작성에 포함되는 변수들은 아래의 <표 4-1>과 같다.

<표 4-1> 단계적방법 Table

Effect Name		Effect Label	Effect Name		Effect Label
1	YK003	재직기간	6	YK026	최대 연체금액
2	YK013	분기수신평잔	7	YK035	보증발생금액
3	YK014	은연 채무불이행 해제건수	8	YK036	신용카드 총 개설건수
4	YK023	현금서비스이용기관수	9	YK045	은행업권 조회건수
5	YK025	단기연체 금액의 합	10	YK046	대부업권 조회건수

4.1.3 변수변환

자료의 정제과정에서 변수변환을 통해 범주화되었고 신용상태에 대한 로지스틱 회귀분석방법을 사용하였기 때문에 신용평점표 작성시에도 변수변환은 범주화에 중점을 두었다. 범주화를 위한 변환은 SAS E-miner의 Interactive grouping node를 사용하였으며 범주화된 변수들도 재범주화하였다. Interactive grouping node는 범주화된 변수의 분포, 기초통계량, 범주화된 기준과 결과를 세부적으로 제시해준다. Interactive grouping node의 Interactive option을 실행한 결과는 아래의 <표 4-2>와 같다.

10) 전진선택법에서 10개, 후방제거법에서 9개, 단계적방법에서 9개의 변수가 최종 선택되었다. 전진선택법에서 은행업권 조회건수가 추가로 식별되었으며 각 방법에서 선별해낸 나머지 9개의 변수는 동일하였다.

<표 4-2> 변수별 범주화 기준

구분		Class	세부기준
재직기간	C_YK003	1	10년미만
	C_YK003	2	10년이상 20년 미만
	C_YK003	3	20년 이상
분기수신 신규	C_YK013	1	100만원 이상
	C_YK013	2	100만원 미만
은연 채무불이행 해제건수	C_YK014	1	없음
	C_YK014	2	1~2건
	C_YK014	3	3건 이상
현금서비스 이용기관수	C_YK023	1	1건
	C_YK023	2	2개 이상
연체건수	C_YK024	1	없음
	C_YK024	2	1회
	C_YK024	3	2~3회
	C_YK024	4	4회 이상
신용카드개설건수	C_YK036	1	1개
	C_YK036	2	2~3개
	C_YK036	3	4~5개
	C_YK036	4	6개 이상
대부업권조회건수	C_YK046	1	있음
	C_YK046	2	없음

4.2 신용평점표 작성

범주화된 변수들의 추정치를 기준으로 가중치를 부여하여 평점을 산출하였다. 추정치들에 POD(Point of Double odds)[11]의 개념을 적용하여 아래의 <표 4-3>에 제시된 보정된 추정치를 기준으로 각 범주별로 평점을 부여하였다. <표 4-2>의 분류기준과 <표 4-3>의 평점을 종합하여 해석해보면 재직기간의 경우 20년 이상된 관측치가 가장 높은 평점을 받고 분기에 100만원 이상 예금성 상품에 가입되어 있는 경우, 은행연합회의 채무불이행 해제건수가 없고 현금서비스를 이용하지 않거나 이용하더라도 1개의 기관을 이용하는 경우 연체와 대부업권 조회건수가 없고 신용카드를 불필요하게 많이 가지고 있지 않은 경우 높은 평점을 받고 있음을 알 수 있다.

11) 평점은 자연로그단위로 산출되며 경우에 따라서 음의 값을 가질 수도 있으므로 POD를 이용하여 선형단위로 변환한다.(Scallan, 1999) 본 연구에서는 POD를 50으로 적용하여 신용평점표를 산출하였다.

<표 4-3> 변수범주별 평점

구분	Class	DF	weighted	score	χ^2	P-value	exp(est)
Intercept		1	5.0764	134	135.71	<.0001	0.006
C_YK003	1	1	0.4055	29	8.46	0.0036	0.667
	2	1	0.227	16	1.91	0.1665	0.836
	3	0	0	0	.	.	.
C_YK013	1	1	0.3403	25	.	.	.
	2	0	0	0	5.14	0.0234	1.405
C_YK014	1	1	1.449	105	48.36	<.0001	4.259
	2	1	0.602	43	4.05	0.044	1.826
	3	0	0	0	.	.	.
C_YK023	1	1	1.232	89	69.84	<.0001	3.428
	2	0	0	0	.	.	.
C_YK024	1	1	1.5163	109	140.36	<.0001	4.555
	2	1	0.795	57	20.66	<.0001	2.214
	3	1	0.6389	46	6.86	0.0088	1.894
	4	0	0	0	.	.	.
C_YK036 [12]	1	1	0.7435	54	11.21	0.0008	2.103
	2	1	0.8832	64	19.83	<.0001	2.419
	3	1	0.5764	42	8.27	0.004	1.78
	4	0	0	0	.	.	.
C_YK046	1	1	2.4052	173	57.85	<.0001	11.081
	2	0	0	0		.	.

4.3 신용평점표 평가

4.2절에서 산출한 신용평점표의 결과해석은 자료나 분석방법에 대한 해박한 지식이 없더라도 이해할 수 있는 평범한 사실이지만 신용평점표가 특정한 기준으로서 인정받기 위해서는 평가를 통해 모형으로서의 타당성을 입증받아야한다. 본 연구에서는 신용평점표의 타당성을 평가하기 위해 K-S 통계량과 ROC, 안정성을 평가하기 위해 PSI를 사용하였다. 결론부터 이야기하자면 타당성과 안정성이 입증되어 신용평점표로서의 충분한 역할을 할 수 있는 것으로 분석되었다.

12) 신용카드개설건수가 과도하게 많은 경우 소위 "돌려막기"현상이 발생하여 오히려 신용평점산출시 불리한 요소로 작용되고 있기 때문에 다른 변수들처럼 Class 증가에 따라 가중평점이 증가하지 않고 오히려 감소하였다.

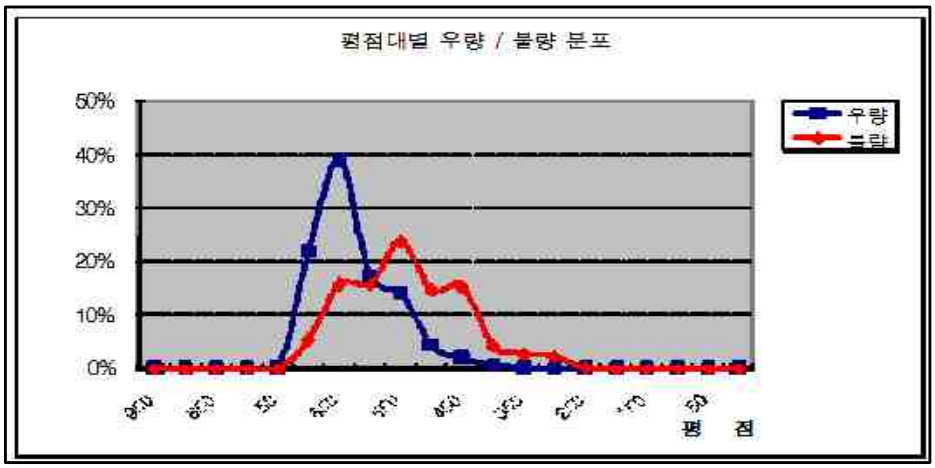
4.3.1 예측성능평가

① K-S 통계량

그림 <4-1>은 평점대별로 신용상태가 우량인 관측치와 불량인 관측치의 분포이다. 우량과 불량의 분포과 확연하게 구분되고 있을뿐만 아니라 우량의 경우 불량보다 평점이 높은 곳에 분포하고 있음을 알 수 있다.

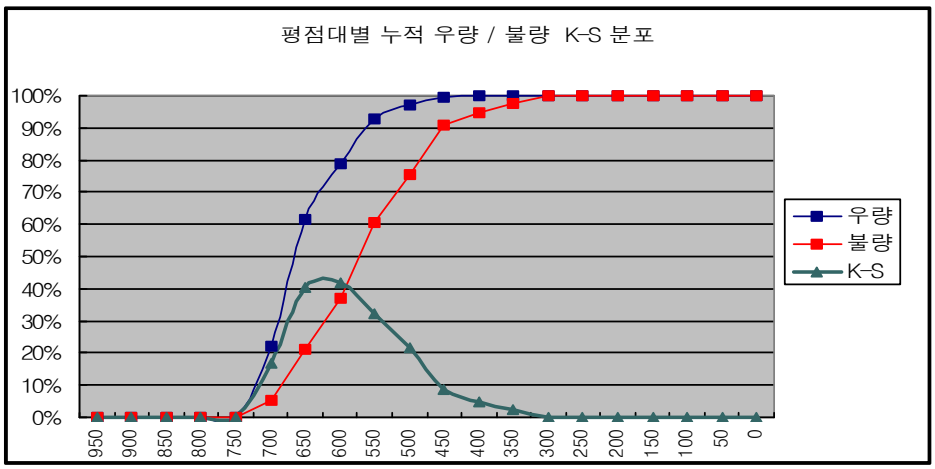
<표 4-4> 평점대별 우량 / 불량 구성 및 K-S통계량

등급	점수대	해당건수			구성비			K-S
		건수	우량	불량	%	우량%	불량%	
1	750	0	0	0	0	0	0	0.0%
2	700	434	403	31	18.2%	22.2%	5.4%	16.7%
3	650	804	714	90	33.7%	39.3%	15.8%	40.2%
4	600	402	312	90	16.9%	17.2%	15.8%	41.6%
5	550	392	257	135	16.4%	14.2%	23.7%	32.0%
6	500	163	80	83	6.8%	4.4%	14.6%	21.9%
7	450	124	37	87	5.2%	2.0%	15.3%	8.6%
8	400	34	10	24	1.4%	0.6%	4.2%	4.9%
9	350	17	2	15	0.6%	0.1%	2.6%	2.4%
10	300	14	1	13	0.6%	0.1%	2.3%	0.2%
11	250	1	0	1	0.0%	0.0%	0.2%	0.0%
	합계	2,385	1,816	569	99.9%	100.0%	100.0%	41.6%



<그림 4-1 > 평점대별 우량 / 불량분포

이러한 분포차이 정도를 K-S(Kolomogrov-Smirnow) 통계량을 통해 확인할 수 있다. K-S 통계량은 신용상태가 우량인 관측치와 불량인 관측치가 분리되는 정도를 측정하는 수치로서 누적 우량비율과 누적 불량비율 분포의 최대 차이값을 의미한다 (임종건, 2003). 아래의 <그림 4-2>에 나타나듯이 산출된 신용평점표의 K-S의 최대값은 41.6으로 좋은 성능을 보여주고 있다[13].



<그림 4-2 > 평점대별 누적 우량 / 불량 / K-S 분포

13) K-S 통계량을 판단하는 일반적인 지침은 "20이하 : 이용가치가 희박한", "20~40 : 적당한", "40~50 : 좋은", "50~60 : 매우좋은", "60~75 : 경이로운", "75이상 : 지나치게 좋은(의심할 필요있는)"으로 준용하고 있다.

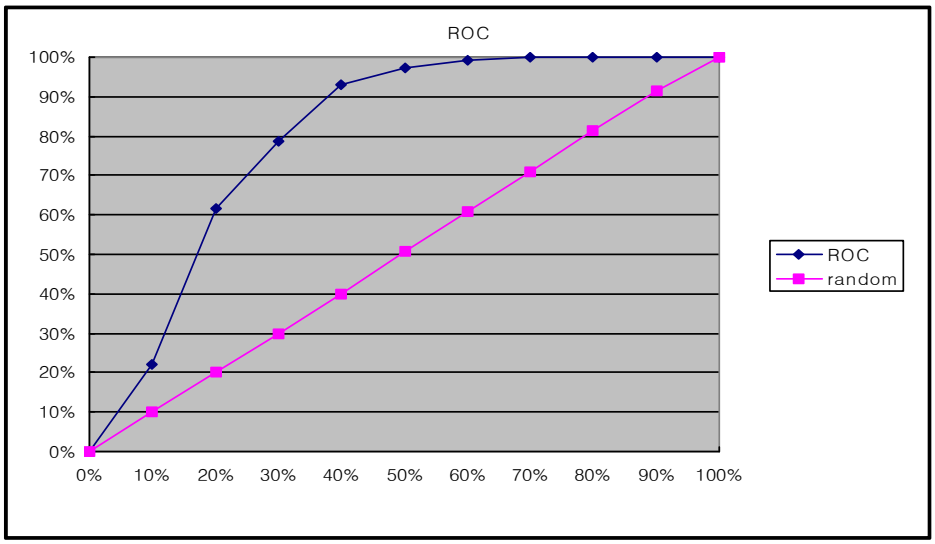
② ROC curve

ROC curve는 판별의 정확도를 도표로 평가하기 위해 전통적으로 사용되어져 온 것으로 제 3장 분석방법론에서도 언급한 바 있다. ROC curve를 작성하기 위해 필요한 민감도와 특이도의 개념을 신용평점표에 적용해 본다면 신용상태를 우량으로 또는 불량으로 얼마나 정확하게 구분해내는 정도를 의미한다고 할 수 있다. 산출된 신용평점표의 ROC curve는 <그림 4-3>와 같고 C-통계량[14] 값은 76.8이므로 예측모형의 성능이 좋은 편임을 알 수 있다.

<표 4-5> C-통계량 산출표

등급	점수대	구성비			누적 구성비			ROC (C-통계량)
		%	우량%	불량%	%	우량%	불량%	
1	750	0	0	0	0	0	0	0.0%
2	700	18.2%	22.2%	5.4%	18.2%	22.2%	5.4%	0.6%
3	650	33.7%	39.3%	15.8%	51.9%	61.5%	21.3%	6.6%
4	600	16.9%	17.2%	15.8%	68.8%	78.7%	37.1%	11.1%
5	550	16.4%	14.2%	23.7%	85.2%	92.8%	60.8%	20.3%
6	500	6.8%	4.4%	14.6%	92.0%	97.2%	75.4%	13.9%
7	450	5.2%	2.0%	15.3%	97.2%	99.3%	90.7%	15.0%
8	400	1.4%	0.6%	4.2%	98.7%	99.8%	94.9%	4.2%
9	350	0.6%	0.1%	2.6%	99.2%	99.9%	97.5%	2.6%
10	300	0.6%	0.1%	2.3%	99.8%	100%	99.8%	2.3%
11	250	0.0%	0.0%	0.2%	99.9%	100%	100%	0.2%
	합계	99.9%	100%	100%	100%	100%	100%	76.8%

14) C-통계량에 근거한 모형의 예측력 판단 기준은 다음과 같다.
"0.5 이하 : 변별력이 없는 것으로 판단", "0.5~0.65 : 변별력이 낮음", "0.65~0.75 : 변별력 보통", "0.75~0.85 : 변별력이 우수함", "0.85~1 : 변별력이 매우 우수함".



<그림 4-3 > 평점표의 ROC curve

4.3.2 안정성평가

PSI (Population Stability Index)[15]

K-S 통계량과 ROC curve를 통해 모형의 타당성이 입증되었다면 분포의 안정성을 확인해볼 필요가 있다. 모집단의 안정성을 측정하기 위한 방법으로 다른 시점에서 분포가 얼마나 변하였는가를 검증하는 PSI를 사용하였다. PSI는 신용평가 모형 개발 전후의 등급별 구성비 차이에 대한 변화정도를 측정한 편차지수라고 할 수 있다. PSI는 데이터의 크기와는 무관하고 시간의 흐름에 따른 분포의 변화를 계량화할 수 있는 지표이다. (임종건, 2006)

본 연구에서는 개발시 등급별 구성비[16]와 검증시 등급별 구성비[17]를 비교하여 전체적인 모형의 안정성을 확인하였다. <표 4-5>에서 산출된 PSI 값은 0.0058[18]이며

15) 개발용 데이터세트와 검증용 데이터세트의 전체적인 분포차이를 확인하기 다음의 수식에 따라 PSI를 산출하였다(김지환, 2007).
$PSI_i = (T_i - V_i) \times \ln(T_i / V_i)$
[T_i: *train* 데이터 세트의 i번째 등급 구성비, V_i: *validation* 데이터 세트의 i번째 등급 구성비]

16) 데이터 분할시 Train 영역으로 분할된 데이터 세트의 우량, 불량 구성비

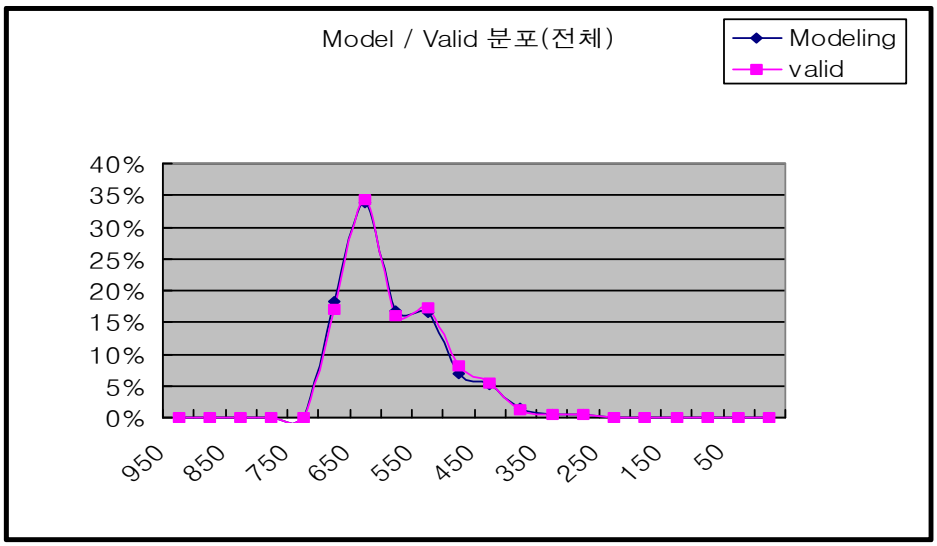
17) 데이터 분할시 Validation으로 분할된 데이터 세트의 우량, 불량 구성비

18) PSI를 기준으로 안정성을 판단하는 기준은 다음과 같다(은행연합회 CSS 해외연수 보고서, 2000).
"0.10이하 : weak(변화가 거의 없는)", "0.10~0.30 : medium(구성비가 다소변화함),
"0.30 이상 : strong (구성비가 현저하게 변화함)"

이는 개발 모형과 검증 모형의 차이가 작다는 것을 의미한다. 개발 모형과 검증 모형의 전체적인 분포를 비교한 <그림 4-4>에서도 두 모형의 차이가 거의 없음을 확인할 수 있다. 따라서 구축된 신용평점표는 전체적인 분포의 안정성을 확보하고 있다고 할 수 있다.

<표 4-6> 평점별 PSI

평점		PSI
1	700	0.00073
2	650	0.00009
3	600	0.000507
4	550	0.00035
5	500	0.002189
6	450	0.000122
7	400	0.000743
8	350	0.000562
9	300	0.000562
계		0.0058



<그림 4-4> Traning 모형과 Valid 모형의 분포비교

제 5 장 결 론

데이터마이닝을 활용하여 자료의 정제, 신용평가 모형의 구축, 평점표 작성까지의 전반적인 과정에 대해 연구하였다. 본 연구의 주요 결과를 요약하면 다음과 같다.

첫째, 자료 정제의 중요성과 효용성이다. 연속형 변수와 범주형 변수가 혼합되어 있고 결측치가 상당히 많은 원자료에서 자료의 분포를 관찰하고 특이치와 결측치를 제거하여 적절한 형태로 변환시켜 모형구현에 적합한 데이터세트를 확보하였다. 변수를 변환하거나 결측치를 제거하는 과정에서 일부의 정보손실이 있을 수도 있지만 오히려 관리와 분석 및 처리가 용이하고 목표변수를 예측하는데 더욱 가치 있는 자료로서의 역할을 할 수 있다는 것을 확인하였다.

둘째, 의사결정나무분석 방법에 대하여 3가지의 알고리즘을 적용하고 로지스틱 회귀분석방법에서 3가지 변수선택방법을 적용하여 각 방법별 최적의 모형을 도출하였다. 의사결정나무분석에서는 Gini Index를 활용하는 방법이 정확도와 예측성능이 가장 우수하였으며 로지스틱 회귀분석에서는 단계적 방법의 예측성능이 가장 우수하였다. 선별된 각 분석방법론을 비교하였을때에는 의사결정나무분석보다 로지스틱 회귀분석의 예측성능이 비교적 우수하게 나타났다.

셋째, 로지스틱 회귀분석방법을 활용하여 신용평점표를 산출하였다. 신용평점표 산출의 과정에서 변환된 데이터세트중 유의한 변수를 다시 선별하여 재범주화 하였으며 POD를 활용해 회귀계수를 보정하고 평점화하였다. 산출된 평점표는 K-S 통계량과 ROC의 C-통계량을 통해 예측성능을 평가하였고 PSI를 활용하여 안정성을 평가하여 모형이 적절하게 구축되었는지 확인하였다.

판단미정의 관측치가 전체적인 분석결과에 큰 영향을 주지는 않았으나 연구 진행에서 제외시켰던 것과 변수를 변환하고 제거하는 과정에서 발생하는 정보의 손실과 이를 최소화하기 위한 방법에 대한 부분을 포함하여 연구를 진행하였다면 더욱 정확한 연구결과와 심도있는 사실을 발견할 수도 있었겠지만 오히려 복잡하고 어려운 모형보다 기본적인 관점에서 접근하여 쉽게 적용할 수 있고 이해할 수 있는 결과를 타당성 있게 제시할 수 있었다는 것이 본 연구의 가장

큰 성과라고 할 수 있다. 신용평가시스템은 날이 갈수록 예측성능이 진화하고 정교해지고 있으며 분석도구와 방법론도 다양해지고 있으나 변수변환과 분석방법론을 적용해 최종 모형을 선택하고 평점표를 산출하는 기본적인 골격은 큰 변화가 없다. 신경망 분석과 복합모형등 다른 방법론을 적용하고 비교분석하였더라면 더욱 양질의 연구성과가 있었을 수도 있겠지만 추후 연구과제로 남기고 본 연구를 마치고자한다.

참 고 문 헌

강현철, 한상태, 최종후, 이성건, 김은석, 엄익현, 김미경 (2007).

고객관리(CRM)를 위한 데이터마이닝 방법론, 자유아카데미, 서울

김지환 (2007) 신용평가모형의 적합성 검증, 연세대학교 대학원 석사학위논문, 서울

김민정 (2001) Credit Risk Modeling Design and Application "신용위험평가모형 이론과 실제",

넥스트웨이브, 서울

이명식, 김정인 (2007) 개인신용평점제도 이론과 실제, 서울출판미디어, 서울

임종건 (2003) 신용평가시스템 적합성검증, Risk Review, 금융감독원

임종건 (2006) 신용등급 계량화에 대한 적합성 검증 방법론, Risk Review, 금융감독원

최종후, 한상태, 강현철, 김은석, 김미경, 이성건 (2001) SAS E-miner 4.0을

이용한 데이터마이닝, 자유아카데미, 서울

최종후, 소선하 (2005) 사례로 배우는 데이터마이닝, 자유아카데미, 서울

CSS 실무위원회 (2000) Credit Scoring System 해외연수보고서, 전국은행연합회

Abstract

Study for Credit Score modeling through data mining methodology

This research is for the comprehension of general procedures in credit assessment modeling and scoring system. SAS E-miner is applied over the whole procedures including Obtaining raw data set, processing and modeling with it. This study consists of 3 section.

First, Comprehension for the raw data set. This process is not only the basic course of this reseach but the most important one. Surveying and Observing distributions of raw data set, appropriate categorical process shows the fact that unmodelized raw data can be modelized and give us more information even if it gives up its own and unique feature.

Secondly specific modelizing procedure. Decision Tree and Regression Methodology are used in this study. In Decision tree analysis method, a model is embodied on three basis algorisms and evaluated through correct classifcation, Lift Graph and ROC curve. The final Decision Tree model is selected by the tree former criterior and it can discern various variables which effect to credit status. In Regression methodology, Logistic Regression method is applied and three variable selection methodologies are used. Each variable selection method is rated by the assessment criterior Lift Graph, ROC curve and others statistics.

According to the result of Comparision with these two methodologies (Lift Graph, ROC curve), the finest model is sorted. And it is Logisitic regression method.

Thirdly procedure for completing credit score table. In this procedure, categorically transformed variables are re-transformed. In each variable

has its own categories and weighted by scores. Credit scores table is generalized on a basis with the weighted scores. K-S statistics and C-statistics estimate the credit score table in a point of suitability view and PSI estiamtes its stability and shift.

Advansed treatment for missing values and complicate methodologies in variable selection(correlation analysis, cluster analysis, neural network) are not applied in this study but it enough to demonstrate the importance of data cleaning procedure and give much information about the whole process of medeling credit scoring system.

부 록

1. 변수변환

<부록 1-1> 변수변환에 따른 변수별 영향력

구분	LABEL	주요통계량			
		p-value	OR	OR Interval	
YK001	차량가격	0.5265	1.017	0.965	1.072
YK002	대출신청금액	0.001	1	1	1
YK003	사업년수	<.0001	0.67	0.576	0.78
YK004	성별(1:남,2:여)	0.1137	1.207	0.956	1.525
YK005	연령	0.0008	0.981	0.97	0.992
YK006	최초 상담일로부터 경과일수	0.1745	0.964	0.914	1.016
YK007	최근 상담일로부터 경과일수	0.0016	0.879	0.811	0.952
YK008	총자산	<.0001	1	1	1
YK009	순자산(자본총계)	0.8115	1	1	1
YK010	납입자본금	0.011	1	1	1
YK011	매출액	0.0038	1	1	1
YK012	영업이익	0.9009	1	1	1
YK013	당기순이익	0.0076	1	1	1
YK014	은연 채무불이행 해제건수	<.0001	1.523	1.36	1.705
YK015	은연 채무불이행 최근 해제일로부터의 기간	<.0001	1.382	1.297	1.473
YK016	총이용금액합계	0.3539	1	1	1
YK017	일시불이용금액합계	0.3481	1	1	1
YK018	총한도합계금액	<.0001	1	1	1
YK019	총이용기관수	0.5011	1.029	0.948	1.116
YK020	총이용잔액합계	0.0666	1	1	1
YK021	할부이용금액합계	0.7574	1	1	1
YK022	현금서비스이용금액합계	0.0149	1	1	1
YK023	현금서비스이용기관수	<.0001	1.957	1.634	2.344
YK024	[2Y] 단기연체 건수	<.0001	1.13	1.104	1.157
YK025	[2Y] 단기연체 금액의 합	0.1792	1	1	1

<부록 1-1> 변수변환에 따른 변수별 영향력

구분	LABEL	주요통계량			
		p-value	OR	OR Interval	
YK026	[2Y] 최대 연체금액	0.4388	1	1	1
YK027	[2Y] 최대 연체경험일수	0.1836	1	1	1
YK028	최초 신용카드 개설일로부터 경과일수	0.0005	0.867	0.8	0.939
YK029	최근 신용카드 개설일로부터 경과일수	0.6998	1.017	0.934	1.108
YK030	현재 보증건수	0.0004	1.084	1.037	1.135
YK031	현재 보증금액의 합	<.0001	1	1	1
YK032	최초 보증 발생일로부터의 경과일수	0.0004	0.84	0.762	0.926
YK033	최근 보증 해지일로부터의 경과일수	0.3861	1	0.999	1
YK034	[3Y] 보증발생건수	<.0001	1.096	1.051	1.142
YK035	[3Y] 보증발생금액	0.0015	1	1	1
YK036	신용카드 총 개설건수	<.0001	0.899	0.86	0.94
YK037	총 신용 개설건수	<.0001	0.911	0.876	0.948
YK038	최근 대출일자로부터 경과일수	0.687	1.012	0.955	1.072
YK039	총 대출 건수	0.3121	1.039	0.965	1.118
YK040	총 대출 금액	0.9893	1	1	1
YK041	최근 조회일로부터 경과일수	<.0001	0.749	0.694	0.808
YK042	최초 조회일로부터 경과일수	0.0162	1.119	1.021	1.226
YK043	전체 조회건수	<.0001	1.089	1.071	1.107
YK044	전체 조회업체수	<.0001	1.224	1.18	1.27
YK045	은행업권 조회건수	0.3851	0.984	0.947	1.021
YK046	대부업권 조회건수	<.0001	6.671	3.94	11.29
YK047	총 조회건수(은행)	0.4342	0.986	0.953	1.021
YK048	총 조회건수(상호저축)	0.1106	1.235	0.953	1.602
YK049	총 조회건수(카드)	<.0001	1.194	1.14	1.251
YK050	총 조회건수(캐피탈)	<.0001	1.064	1.048	1.081

2. 의사결정나무 분석방법

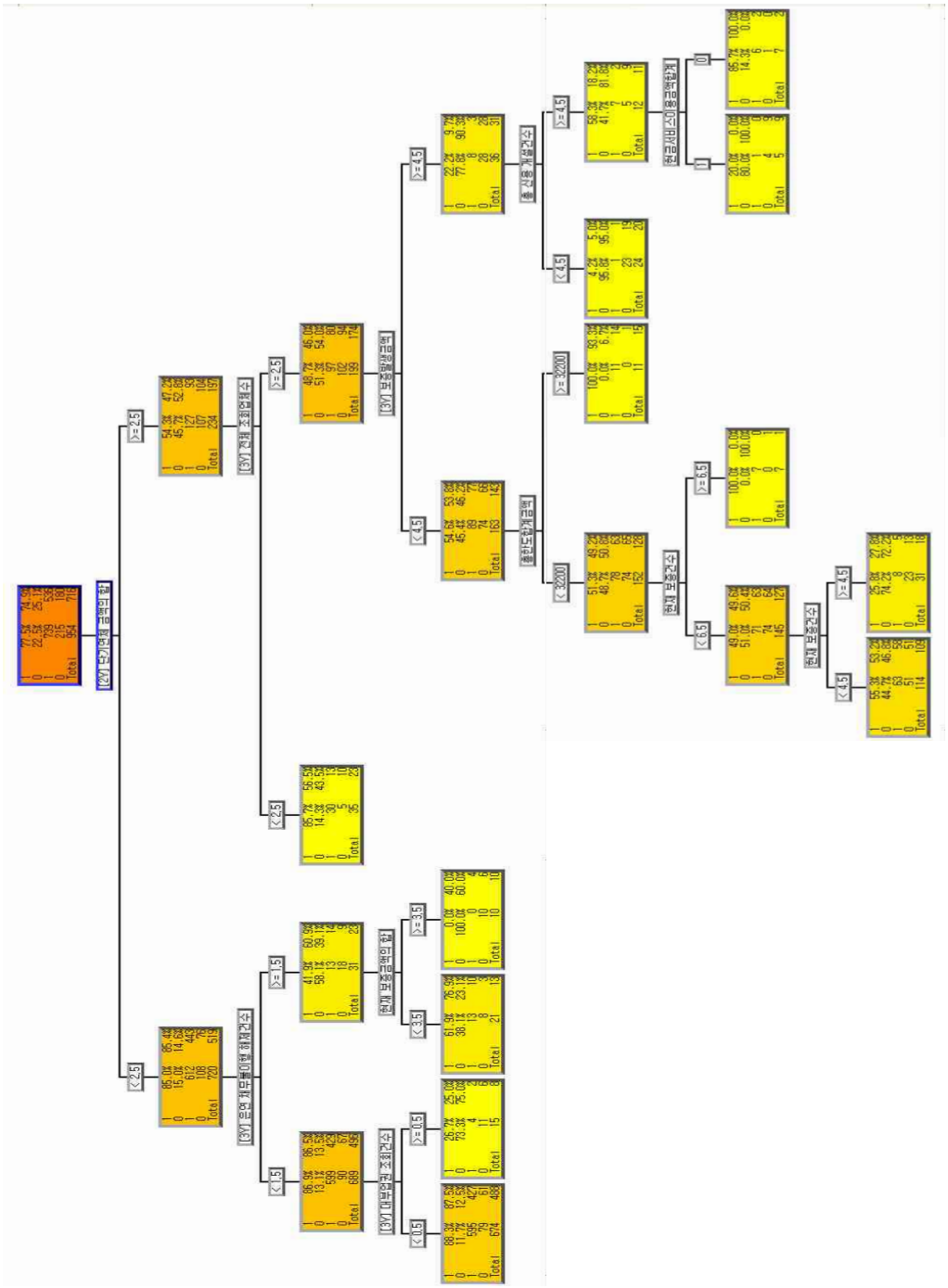
<부록 2-1> 알고리즘별 정오분류율 (단위 : %)

절단값	CHAID				Entrophy				Gini			
	불량		우량		불량		우량		불량		우량	
	정	오	정	오	정	오	정	오	정	오	정	오
0	0	100	100	0	0	100	100	0	0	100	100	0
5	7.22	92.78	99.06	0.94	13.88	86.12	99.06	0.94	6.11	93.89	99.06	0.94
10												
15									12.22	87.78	98.88	1.12
20												
25	8.33	91.67	98.88	1.12	18.88	81.12			13.33	86.67		
30					29.44	70.56	97.76	2.24			98.69	1.31
35	30	70	95.89	4.11					35	65	95.70	4.3
40												
45									45.55	54.45	93.84	6.16
50												
55												
60					50.77	49.23	86.94	13.06	49.44	50.56	91.79	8.21
65	72.22	27.78	79.10	20.90	59.44	40.56	85.07	14.93				
70									67.22	32.28	82.08	17.92
75												
80												
85									71.11	18.89	75.74	24.26
90	98.88	1.12	9.70	90.30	90.88	9.12	2.61	97.39	79.44	10.56	72.20	27.80
95									98.88	1.12	9.70	90.3
100	100	0	0	100					100	0	0	100

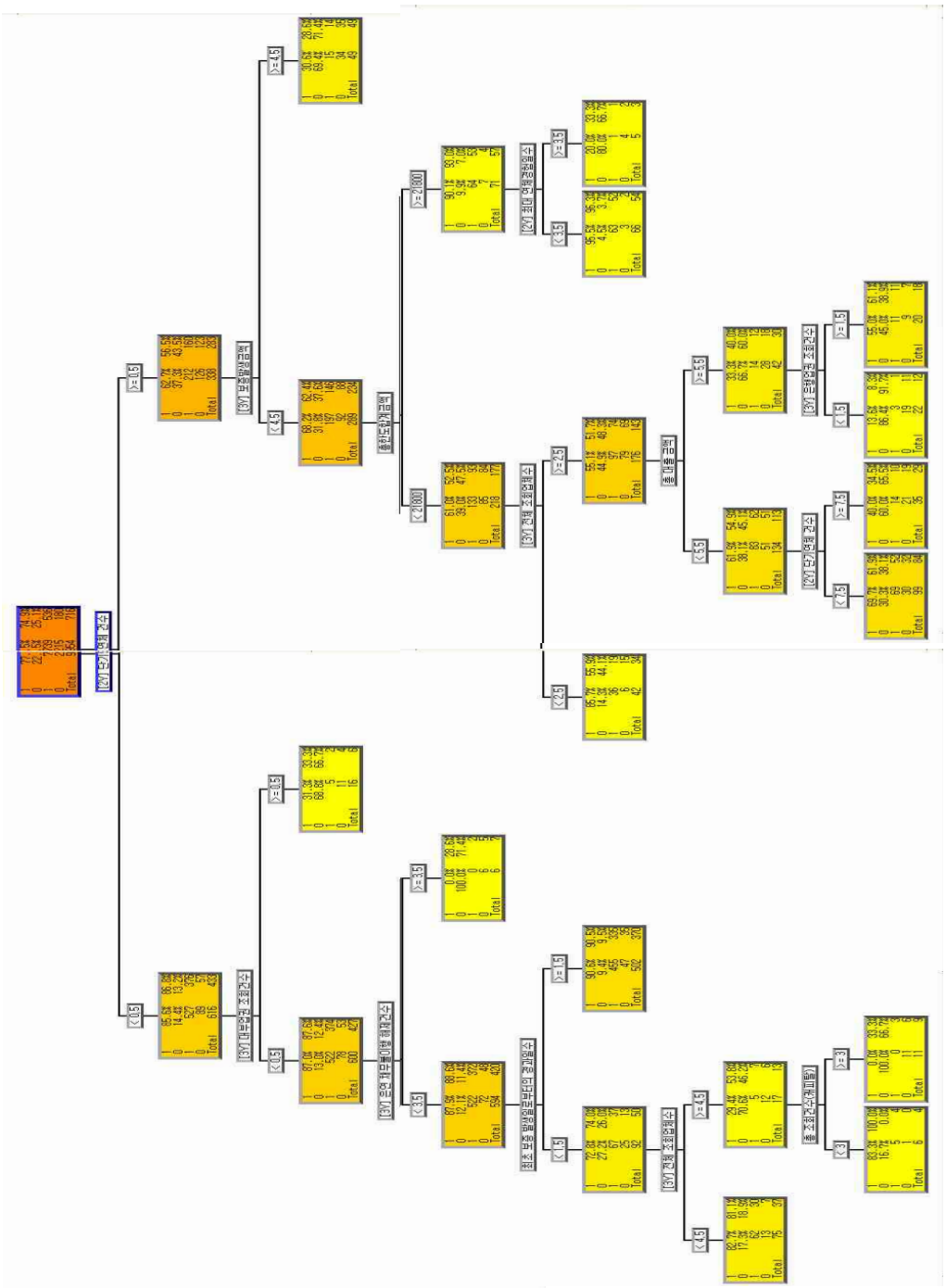
<부록 2-2> 알고리즘별 정확도, 민감도, 특이도

절단값	CHAID			Entrophy			Gini		
	정확도	민감도	특이도	정확도	민감도	특이도	정확도	민감도	특이도
0	74.8	1.00	0.00	74.8	1.00	0.00	74.8	1.00	0.00
5		0.99	0.07		0.99	0.14	75.69	0.99	0.06
10	75.97	0.99	0.07	77.65	0.99	0.14		0.99	0.06
15		0.99	0.07		0.99	0.14	77.23	0.99	0.12
20		0.99	0.07		0.99	0.14		0.99	0.12
25	76.11	0.99	0.08	78.91	0.99	0.19		0.99	0.13
30		0.99	0.08		0.98	0.29		0.99	0.13
35		0.96	0.30		0.98	0.29	80.44	0.96	0.35
40		0.96	0.30	80.58	0.98	0.29			
45	79.32	0.96	0.30		0.98	0.29	81.70	0.94	0.46
50		0.96	0.30		0.98	0.29		0.94	0.46
55		0.96	0.30		0.98	0.29		0.94	0.46
60		0.96	0.30	79.60	0.87	0.51	81.14	0.92	0.49
65		0.79	0.72		0.85	0.59		0.92	0.49
70		0.7	0.72		0.85	0.59		0.82	0.67
75	77.37	0.79	0.72	78.63	0.85	0.59	78.35	0.82	0.67
80		0.79	0.72		0.85	0.59		0.82	0.67
85		0.79	0.72		0.85	0.59	74.58	0.76	0.71
90	32.12	0.10	0.99	26.81	0.03	0.91	74.02	0.72	0.79
95		0.10	0.99		0.03	0.91	32.12	0.10	0.99
100	25.14	0.00	1.00		0.03	0.91	25.13	0.00	1.00

<부록 2-3> Chi-square 알고리즘 의사결정나무 모형



<부록 2-4> Entropy Index 알고리즘 의사결정나무 모형



3.로지스틱 회귀분석방법

<부록 3-1 > 절단값에 따른 변수선택별 정오분류율

(단위 : %)

절단값	전진선택법				후방제거법				단계적방법			
	불량		우량		불량		우량		불량		우량	
	정	오	정	오	정	오	정	오	정	오	정	오
0	536	0	0	180	536	0	0	180	536	0	0	180
5	536	5	0	175	536	5	0	175	536	5	0	175
10	534	9	2	171	534	14	2	166	534	9	2	171
15	533	15	3	165	533	14	3	166	533	15	3	165
20	533	23	3	157	533	19	3	161	532	23	4	157
25	531	29	5	151	532	27	4	153	532	31	4	149
30	530	35	6	145	531	33	5	147	531	36	5	144
35	529	39	7	141	528	37	8	143	529	44	7	136
40	522	47	14	133	526	49	10	131	523	48	13	132
45	518	52	18	128	515	55	21	125	512	52	24	128
50	506	65	30	115	502	65	34	115	508	63	28	117
55	503	78	33	102	486	71	50	109	498	75	38	105
60	484	96	52	84	468	86	68	94	481	77	55	103
65	461	105	75	75	449	97	87	83	462	106	74	74
70	436	120	100	60	426	119	110	61	432	121	104	59
75	400	132	136	48	400	133	136	47	399	134	137	46
80	336	144	200	36	337	147	199	33	355	146	181	34
85	262	154	274	26	253	158	283	22	269	155	267	25
90	170	169	366	11	164	167	372	13	172	168	363	13
95	63	177	473	3	55	178	481	2	62	179	474	1
100	0	180	536	0	0	180	536	0	0	180	536	0

<부록 3-2> 변수선택방법별 정확도, 민감도, 특이도

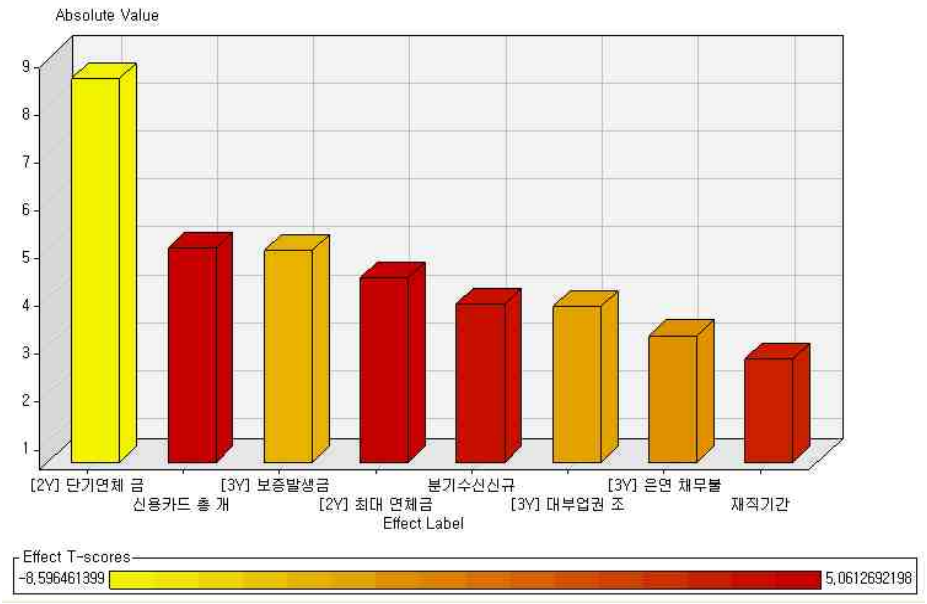
절단값	전진선택법			후방제거법			단계적방법		
	정확도	민감도	특이도	정확도	민감도	특이도	정확도	민감도	특이도
0	0.749	1.000	0.000	0.749	1.000	0.000	0.749	1.000	0.000
5	0.756	1.000	0.028	0.756	1.000	0.028	0.756	1.000	0.028
10	0.758	0.996	0.050	0.765	0.996	0.078	0.758	0.996	0.050
15	0.765	0.994	0.083	0.764	0.994	0.078	0.765	0.994	0.083
20	0.777	0.994	0.128	0.771	0.994	0.106	0.775	0.993	0.128
25	0.782	0.991	0.161	0.781	0.993	0.150	0.786	0.993	0.172
30	0.789	0.989	0.194	0.788	0.991	0.183	0.792	0.991	0.200
35	0.793	0.987	0.217	0.789	0.985	0.206	0.800	0.987	0.244
40	0.795	0.974	0.261	0.803	0.981	0.272	0.797	0.976	0.267
45	0.796	0.966	0.289	0.796	0.961	0.306	0.788	0.955	0.289
50	0.797	0.944	0.361	0.792	0.937	0.361	0.797	0.948	0.350
55	0.811	0.938	0.433	0.778	0.907	0.394	0.800	0.929	0.417
60	0.810	0.903	0.533	0.774	0.873	0.478	0.779	0.897	0.428
65	0.791	0.860	0.583	0.763	0.838	0.539	0.793	0.862	0.589
70	0.777	0.813	0.667	0.761	0.795	0.661	0.772	0.806	0.672
75	0.743	0.746	0.733	0.744	0.746	0.739	0.744	0.744	0.744
80	0.670	0.627	0.800	0.676	0.629	0.817	0.700	0.662	0.811
85	0.581	0.489	0.856	0.574	0.472	0.878	0.592	0.502	0.861
90	0.473	0.317	0.939	0.462	0.306	0.928	0.475	0.321	0.933
95	0.335	0.118	0.983	0.325	0.103	0.989	0.337	0.116	0.994
100	0.251	0.000	1.000	0.251	0.000	1.000	0.251	0.000	1.000

<부록 3-3> 변수선택방법별 C-statistics

전진선택법			후방제거법			단계적방법		
민감도	1-특이도	C	민감도	1-특이도	C	민감도	1-특이도	C
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.118	0.017	0.001	0.103	0.011	0.001	0.116	0.006	0.000
0.317	0.061	0.010	0.306	0.072	0.012	0.321	0.067	0.013
0.489	0.144	0.034	0.472	0.122	0.019	0.502	0.139	0.030
0.627	0.200	0.031	0.629	0.183	0.034	0.662	0.189	0.029
0.746	0.267	0.046	0.746	0.261	0.053	0.744	0.256	0.047
0.813	0.333	0.052	0.795	0.339	0.060	0.806	0.328	0.056
0.860	0.417	0.070	0.838	0.461	0.100	0.862	0.411	0.069
0.903	0.467	0.044	0.873	0.522	0.052	0.897	0.572	0.142
0.938	0.567	0.092	0.907	0.606	0.074	0.929	0.583	0.010
0.944	0.639	0.068	0.937	0.639	0.031	0.948	0.650	0.063
0.966	0.711	0.069	0.961	0.694	0.053	0.955	0.711	0.058
0.974	0.739	0.027	0.981	0.728	0.032	0.976	0.733	0.021
0.987	0.783	0.044	0.985	0.794	0.066	0.987	0.756	0.022
0.989	0.806	0.022	0.991	0.817	0.022	0.991	0.800	0.044
0.991	0.839	0.033	0.993	0.850	0.033	0.993	0.828	0.028
0.994	0.917	0.077	0.994	0.922	0.072	0.993	0.872	0.044
0.994	0.872	-0.044	0.994	0.894	-0.028	0.994	0.917	0.044
0.996	0.950	0.077	0.996	0.922	0.028	0.996	0.950	0.033
1.000	1.000	0.050	1.000	1.000	0.078	1.000	0.972	0.022
1.000	0.972	-0.028	1.000	0.972	-0.028	1.000	1.000	0.028
	계	0.774		계	0.764		계	0.804

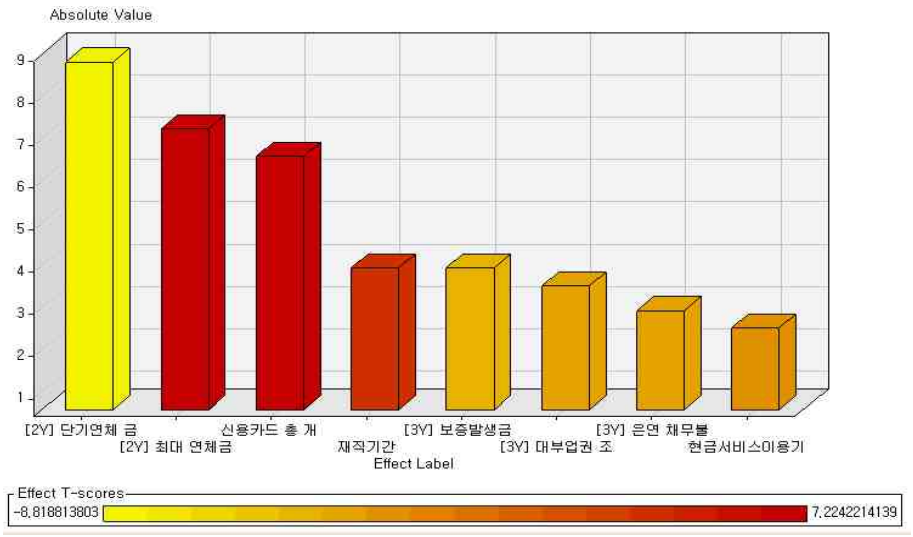
<부록 3-4> 전진선택법의 모수추정치와 T-score

EffectName		EffectLabel	Parameter Estimate	Effect T-scores	Effect Sign
X1	YK003	재직기간	0.0464	2.7307	+
X2	YK013	분기수신신규	0.5835	3.8773	+
X3	YK014	은연채무불이행해제건수	-0.3212	-3.2239	-
X4	YK023	현금서비스이용기관수	-0.4585	-2.5044	-
X5	YK025	단기연체금액의합	-0.5642	-8.5965	-
X6	YK026	최대연체금액	0.4606	4.4278	+
X7	YK035	보증발생금액	-0.3554	-5.0060	-
X8	YK036	신용카드총개설건수	0.2210	5.0613	+
X9	YK045	은행업권조회건수	0.0552	1.4992	+
X10	YK046	대부업권조회건수	-1.9558	-3.8379	-



<부록 3-5> 후방제거법의 모수추정치와 T-score

Effect name		Effect Label	Parameter Estimate	Effect T-scores	Effect Sign
X1	YK003	재직기간	0.0662	3.9430	+
X2	YK014	은연 채무불이행 해제건수	-0.2790	-2.9209	-
X3	YK023	현금서비스이용기관수	-0.4518	-2.5052	-
X4	YK025	단기연체 금액의 합	-0.5750	-8.8188	-
X5	YK026	최대 연체금액	0.6529	7.2242	+
X6	YK035	보증발생금액	-0.2545	-3.9360	-
X7	YK036	신용카드총개설건수	0.2756	6.5894	+
X8	YK045	은행업권 조회건수	0.0625	1.6511	+
X9	YK046	대부업권 조회건수	-1.7945	-3.5184	-



4.신용평점표

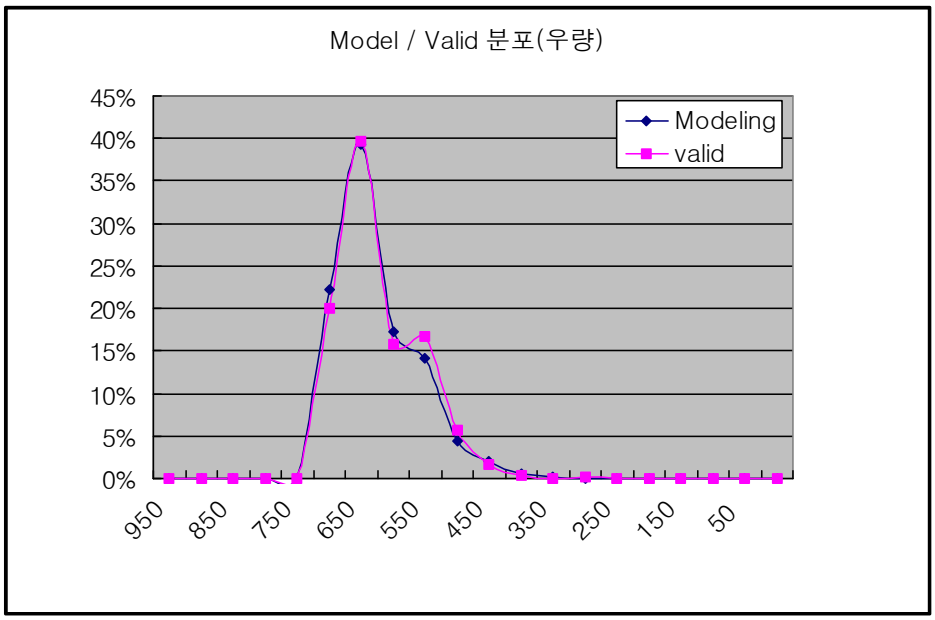
<부록 4-1> 개발 데이터의 구성

등급	점수대	개발 데이터 (해당건수)			개발 데이터 (구성비:%)			개발 데이터 (누적 구성비:%)		
		건수	우량	불량	전체	우량	불량	전체	우량%	불량%
1	750	0	0	0	0	0	0	0	0	0
2	700	434	403	31	18.2%	22.2%	5.4%	18.2%	22.2%	5.4%
3	650	804	714	90	33.7%	39.3%	15.8%	51.9%	61.5%	21.3%
4	600	402	312	90	16.9%	17.2%	15.8%	68.8%	78.7%	37.1%
5	550	392	257	135	16.4%	14.2%	23.7%	85.2%	92.8%	60.8%
6	500	163	80	83	6.8%	4.4%	14.6%	92.0%	97.2%	75.4%
7	450	124	37	87	5.2%	2.0%	15.3%	97.2%	99.3%	90.7%
8	400	34	10	24	1.4%	0.6%	4.2%	98.7%	99.8%	94.9%
9	350	17	2	15	0.6%	0.1%	2.6%	99.2%	99.9%	97.5%
10	300	14	1	13	0.6%	0.1%	2.3%	99.8%	100%	99.8%
11	250	1	0	1	0.0%	0.0%	0.2%	99.9%	100%	100%
12	200	0	0	0	0	0	0	100%	100%	100%
	합계	2,385	1,816	569	99.9%	100%	100%	100%	100%	100%

<부록 4-2> 검증 데이터의 구성

등급	점수대	검증 데이터 (해당건수)			검증 데이터 (구성비:%)			검증 데이터 (누적 구성비:%)		
		건수	우량	불량	전체	우량	불량	전체	우량	불량
1	750	0	0	0	0	0	0	0	0	0
2	700	122	108	14	17.1%	20.0%	8.0%	17.1%	20.0%	8.0%
3	650	245	215	30	34.3%	39.7%	17.2%	51.3%	59.7%	25.3%
4	600	114	85	29	15.9%	15.7%	16.7%	67.3%	75.4%	42.0%
5	550	123	90	33	17.2%	16.6%	19.0%	84.5%	92.1%	60.9%
6	500	58	31	27	8.1%	5.7%	15.5%	92.6%	97.8%	76.4%
7	450	39	9	30	5.5%	1.7%	17.2%	98.0%	99.4%	93.7%
8	400	8	2	6	1.1%	0.4%	3.4%	99.2%	99.8%	97.1%
9	350	3	0	3	0.4%	0.0%	1.7%	99.6%	99.8%	98.9%
10	300	3	1	2	0.4%	0.2%	1.1%	100%	100%	100%
11	250	0	0	0	0.0%	0.0%	0.0%	100%	100%	100%
12	200	0	0	0	0	0	0	100%	100%	100%
	합계	715	541	174	100%	100%	100%	100%	100%	100%

<부록 4-4> 우량의 경우 Model / Valid 분포



<부록 4-5> 불량의 경우 Model / Valid 분포