Journal of the Korean Data & Information Science Society
2013, **24**(3), 411–425

http://dx.doi.org/10.7465/jkdi.2013.24.3.411
한국데이터정보과학회지

분류모형을 이용한 여신회사 고객대출 분석에 관한 연구

김태형[1] · 김영화[2]

[1]SAS Korea · [2]중앙대학교 응용통계학과

접수 2013년 3월 12일, 수정 2013년 4월 17일, 게재확정 2013년 4월 22일

요 약

데이터마이닝이란 대용량의 자료로부터 의미있는 패턴과 규칙을 찾기 위해서 자동화되거나 반자동화된 도구를 이용하여 데이터를 탐색하고 분석하는 과정이다. 이러한 데이터마이닝 기법을 통해 정보의 연관성을 파악함으로써 가치 있는 정보를 만들어 합리적인 의사 결정이 가능하게 된다. 금융분야에서도 데이터베이스 마케팅, 신용평가, 서비스 품질개선, 부정행위 적발 등에 데이터마이닝 기법이 다양하게 사용되고 있다. 금융거래에서 대출의 중요도와 필요성이 시간이 지날수록 점점 높아지고 있으나, 대출을 이용하는 사람과 대출건수가 증가할수록 부실대출의 위험이 함께 증가하기 때문에 대출을 해주는 여신기관의 손실을 막기 위해서는 대출여부를 정확하게 예측할 필요성이 존재한다. 본 연구에서는 국내 A 여신기관의 실제 데이터를 사용하여 대출심사에 관한 연구를 진행하였으며, 모형 구축에 있어서 안정적이고 정확한 예측을 보이는 모형을 찾기 위하여 원 데이터에서의 샘플 정제와 여러 가지 모형, 데이터마이닝 기법 등을 사용하여 다양한 모형을 구축하고 비교, 평가하였다.

주요용어: 대출, 데이터마이닝, 리스크 관리, 빅데이터, 오버샘플링, 의사결정나무.

1. 서론

오늘날 데이터는 점차 방대해지고 있으며 지수적인 증가 속도를 보이고 있다. 시장조사기관 IDC 보고서에 따르면 데이터 정보량이 최근 5년 사이에 180제타 바이트에서 1800제타바이트로 무려 10배나 증가하였다고 한다. 이처럼 데이터가 기하급수적으로 증가함에 따라 사람들은 빅데이터 (big data)의 가치를 새롭게 인식하고 있으며 이를 효과적으로 관리하고 분석하여 최적의 프로세스를 찾으려는 노력이 많이 시도되고 있다. '데이터는 21세기의 원유다.'라는 말이 있듯이 방대한 데이터에서 얻어낼 수 있는 정보는 무한하며 여기서 유의미한 정보를 찾아내었을 때, 정보는 비로소 우리가 필요로 하는 지식으로 전환된다. 한편 개인의 특성이 다양화되고 독특한 개성을 지닌 사람들의 증가로 인하여 데이터 또한 점점 다양화 되어지고 있으며 각 개인의 특성을 정확하게 분석하여 의미있는 결과를 유도해내는 것이 점차 어려워지고 있는 실정이다. 이러한 문제의 해결책으로 제시되어 활발하게 사용되고 있는 것이 데이터마이닝 기법이다. 여기서 데이터마이닝이란 대용량 자료로부터 의미있는 패턴과 규칙을 발견하기 위해서 자동화되거나 반자동화된 도구를 이용하여 데이터를 탐색하고 분석하는 과정이다 (Berry와 Linoff, 1997, 2011). 이러한 데이터마이닝 기법을 통해 정보의 연관성을 파악함으로써 기업에서는 더욱 가치있는 정보를 만들어 의사 결정에 적용함으로써 이익을 극대화시킬 수 있다. 기업이 보유하고 있는 일일 거래 데이터, 고객 데이터, 상품 데이터 또는 각종 마케팅 활동의 고객 반응 데이터 등과 이외의 기

[1] (135-839) 서울특별시 강남구 대치4동 889, SAS Korea, 컨설턴트.
[2] 교신저자: (156-756) 서울특별시 동작구 흑석동 221, 중앙대학교 응용통계학과, 교수.
E-mail: gogators@cau.ac.kr

타 외부 데이터를 포함하는 모든 사용 가능한 근원 데이터를 기반으로 감춰진 지식, 기대하지 못했던 경향 또는 새로운 규칙 등을 발견하고, 이를 실제 비즈니스 의사 결정 등을 위한 정보로 활용하고자 하는 것이다.

금융분야에서 데이터마이닝의 활용으로는 데이터베이스 마케팅, 고객신용평가, 서비스 품질개선, 금융사기 적발 등이 있으며 본 논문에서는 국내의 A 여신기관의 실제 데이터를 사용하여 대출심사에 관한 연구를 진행하였다. 2012년 금융감독원에서는 부실위험이 있는 저축은행 4곳을 선별하여 영업정지를 발표하였다. 이러한 저축은행의 부실의 근원에는 여러 가지가 있겠지만 은행들의 공격적 마케팅 강화와 외국자본 유입으로 인한 저축은행의 경쟁력 약화도 한 몫 했다고 할 수 있다. 이에 더해 저축은행들이 몸집 불리기에 집중한 나머지 리스크 관리에는 상대적으로 소홀하여 불량채권 회수에 대한 최선의 대응책을 마련하지 못하였다. 이처럼 여신기관에서는 리스크 관리 (risk management)가 매우 중요하며 리스크 관리에 소홀하거나 실패하게 되면 이는 바로 영업적 손실로 직결되며 최악의 경우 정상적인 고객들에게 막대한 경제적 손실을 입히게 된다. 따라서 여신기관의 고객에 대한 대출 가능여부 예측과 판단은 금융 전문가의 치밀한 분석을 통하여 관리가 이루어져야 할 것이다. 본 연구에서는 A 여신기관의 실제 자료를 대상으로 데이터마이닝 알고리즘을 사용하여 여러 가지 모형을 구축하고 그 성능을 비교하고 해당 데이터에 가장 잘 적합하는 모형을 SAS의 E-miner에서 제안하는 분석의 5단계인 SEMMA를 고려하여 A 여신기관의 실제자료에 대해 모형을 생성하고 비교, 평가하였다.

2. 데이터 사전처리

현실세계에서 접하는 데이터들은 논리적으로 정리되어 있지 않거나 변수간 일관성이 없는 경우가 많다. 이러한 결과의 원인으로는 관측자의 주관적인 편향과 입력자의 코딩오류, 측정도구 및 고객들의 응답 기입 실수 등 매우 다양하다. 이렇게 가공되지 않은 원 데이터 (raw data)로 분석을 진행하게되면 고도로 숙련된 분석전문가가 분석을 하더라도 그 결과를 신뢰할 수 없으며 예측 및 분류의 성능도 많이 낮아지는 문제를 야기한다. 따라서 신뢰성을 갖춘 분석을 하기 위해서는 데이터의 사전처리가 매우 중요하다. 본 연구에서 다루고 있는 신용대출과 카드대출의 자료 또한 신뢰성있는 분석결과를 도출하기 위해 원 데이터의 이상치를 탐지, 제거하고 결측값을 보간하는 사전처리 과정을 거쳐 분석하였다.

2.1. 자료 설명

본 연구에서는 카드대출 491,105건, 신용대출 159,580건의 두 개의 자료를 사용하여 실증분석을 실시하였다. Table 2.1은 카드대출에 관한 자료의 설명이다. 변수는 크게 인적사항 변수와 대출관련 변수 그리고 목표변수 3가지로 구성되어 있다. 인적사항 변수로는 직업, 성별, 거주지역, 연령이 있으며, 대출관련 변수로는 범주형 자료인 고객등급, 연속형자료인 대출정보 변수가 있다. 여기서 목표변수는 카드대출 신규이용여부이다.

Table 2.1 Data description of card loan

Variable	Characteristic/Items
Related to card loan	5 variables (continuous type)
Related to service	3 variables (continuous type)
Customer level	Level 1 ∼ Level 4
Gender	Male/Female
Age	20∼29/30∼39/40∼49/50∼59/60∼
Residence	Area 1 ∼ Area 10
Occupation	Job 1 ∼ Job 7
New card loan borrower	Target variable (Yes/No)

Table 2.2는 신용대출에 관한 자료의 설명이며, 카드대출과 마찬가지로 변수를 인적사항 변수, 대출

관련 변수, 목표변수 총 3가지로 구분할 수 있으며 인적사항 변수로는 직업, 성별, 거주지역, 연령 등이 있다. 또한 대출관련 변수는 고객위험도, 고객등급 외는 모두 연속형이다. 여기서 목표변수는 신용대출 신규이용여부이다.

Table 2.2 Data description of credit loan

Variable	Characteristic/Items
Related to credit loan	10 variables (continuous type)
Related to credit card	8 variables (continuous type)
Related to service	2 variables (continuous type)
Customer risk level	Level 1 ∼ Level 7
Customer type	A/B/C/D
Customer grade	Level 1 ∼ Level 4
Gender	Male/Female
Age	20∼29/30∼39/40∼49/50∼59/60∼
Residence	Area 1 ∼ Area 10
Occupation	Job 1 ∼ Job 7
New credit loan borrower	Target variable (Yes/No)

2.2. 이상치제거 및 결측값 대체

데이터 탐색과정에서 이상치로 의심되는 값들을 단순히 제거하면 정보 손실의 문제가 발생하므로 이를 방지하기 위하여 이상치를 다른 값으로 대체하여 분석하는 것이 일반적이다. 또한 이상치 뿐만 아니라 데이터 셋 내의 결측치 또한 적절한 값으로 대체해야한다. 왜냐하면 결측치가 하나라도 존재하는 관찰치는 해당 관찰치의 모든 값이 분석에서 제외되기 때문이다 (Kang과 Han, 1999).

이상치가 존재하게 되면 구축된 모형의 분류 및 예측의 성능이 많이 떨어지게 된다. 또한 잘못된 모형이 만들어질 가능성이 높기 때문에 이상치를 제거하거나 그 값을 적절한 값으로 바꾸어 주어야 한다. 이상치를 제거하는 방법으로는 분포를 이용하여 통계적으로 제거하는 방법과 분석자의 판단하에 제거하는 방법이 있다. 통계적으로 이상치를 제거할 때에는 모든 변수에 일괄처리를 가할 수 있어 분석 프로세스가 단축되어 빠른 분석을 할 수 있는 장점이 있지만, 통계이론적인 관점이 항상 정확하다는 보장이 없기 때문에 실제 데이터에 대해서는 분석자의 경험적인 판단 하에 이상치를 제거하는 방법이 더 좋은 결과를 유도할 수도 있다. 본 연구에서는 이상치를 제거할 때 변수 각각의 의미를 파악하고 논리적으로 맞지 않거나 일관성이 없는 변수들의 값을 제거하였다.

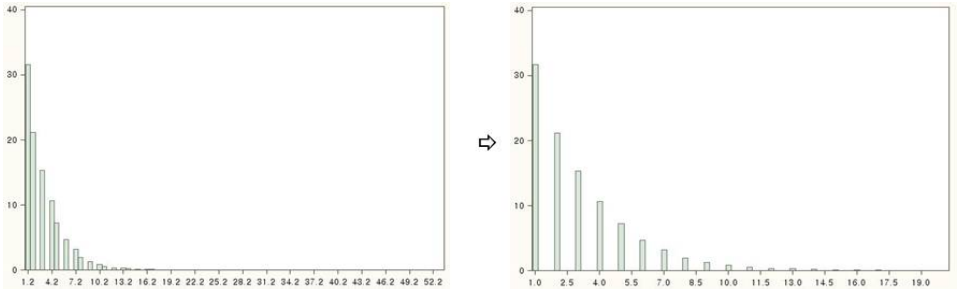

Figure 2.1 Removal effect of outliers in loan variables

Figure 2.1은 이상치를 가지고 있는 변수의 분포를 그려보고 특정값을 기준으로 하여 이상치를 제거한 결과이다. 이상치 제거 전의 분포는 관측치들이 좌측에 많이 존재하였고, 이상치 제거 후의 분포는 관측치들이 우측으로 조금 이동한 것을 볼 수 있다. 본 연구에 사용한 자료들 대부분은 연속형 변수이며 이러한 절차로 이상치를 제거하였다. 이상치를 제거하거나 데이터가 미입수된 경우 결측값이 발생하게 되는데 모형의 성능을 높이기 위해서는 이 결측값을 적절한 값으로 보간해 주어야 한다. 보간을 할 때에

도 통계적 방법과 분석자의 경험적인 판단으로 결측값을 대체하는 두 가지 방법이 있으며, 분석의 간결화를 위해 범주형 자료에서 결측값은 최빈값 (mode)으로 대체하였다. 또한 연속형 변수는 본 연구에서 사용한 자료들의 특성상 이상치가 많기 때문에, 평균 (mean)을 사용하지 않고 중위수 (median)를 사용하여 분석하였다. 한편, 이상치가 무조건 잘못된 정보만을 주는 것이 아닌 경우도 있기 때문에 그럴 경우 분석자의 판단 하에 해당 변수의 최대값으로 값을 대체하였다.

2.3. 오버샘플링

오버샘플링 (oversampling)이란 데이터마이닝 과정 중 모집단으로부터 결과들의 비율을 조절하기 위해 결과들이 드물게 발생하는 사례가 되는 집단으로부터는 많은 수의 표본을 추출하고, 이에 대조가 되는 집단으로 부터는 적은 비율로 표본을 추출하여, 분석을 위한 새로운 모형의 집합을 생성하는 과정이다 (Chung 등, 2008).

앞서 설명하였듯이 목표변수의 수준이 두 개이고 사례 집단이 표본에서 차지하는 비율이 아주 미미하게 추출되었을 때 예측 모형을 구축하는데 어려움이 있으며 신뢰성이 없는 결과를 도출하게 된다. Berry와 Linoff (1997, 2011) 등 많은 선행연구 등에서 사례 집단과 대조 집단의 비율이 대략적으로 1:4 정도로 구성되어질 때 분류 및 예측의 성능이 좋아진다고 밝히고 있다. 본 연구에서 사용하는 신용대출과 카드대출 자료의 경우, 신용대출은 목표변수의 비율이 16:84이며, 카드대출은 목표변수의 비율이 1:99이다. 따라서 신용대출은 오버샘플링 없이 분석을 진행하였고, 반면에 카드대출은 목표변수의 사례 집단과 대조 집단의 비율 차이가 매우 크기 때문에 오버샘플링을 실시하여 분석을 진행하였다.

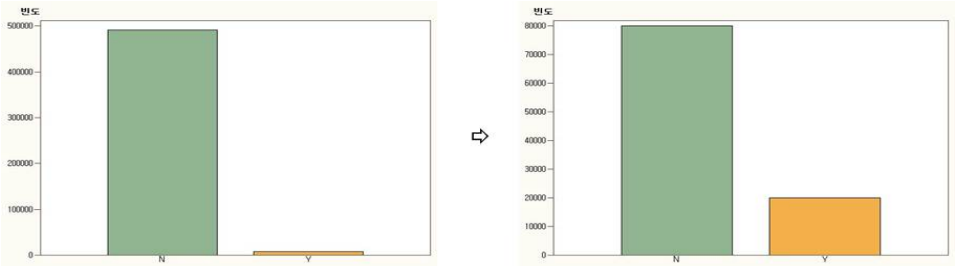

Figure 2.2 Oversampling result of asymmetric data

Figure 2.2는 목표변수의 오버샘플링 전과 후의 분포이다. 오버샘플링 전의 목표변수 N의 관측치 수는 약 49만개 이고, Y의 관측치 수는 7천개로 약 1:99의 비대칭적인 비율을 가지고 있었다. 여기서 목표변수의 비율을 1:4로 오버샘플링 하기 위해 N은 49만개의 관측치 중 8만개를 비복원 무작위 표본추출을 하였고, Y는 7천개의 관측치 중 2만개를 중복을 허용하고 무작위 표본추출을 하여 목표변수의 사례 집단과 대조 집단의 비율을 1:4로 재구성하였다. 다음 Figure 2.3은 EG (Enterprise Guide 4.3)를 사용하여 원 데이터에서 오버샘플링을 하는 절차로서, 목표변수에서 각각의 수준별로 필터링하여 무작위 샘플링을 한 후에 다시 데이터를 결합하는 과정을 보여주고 있다.

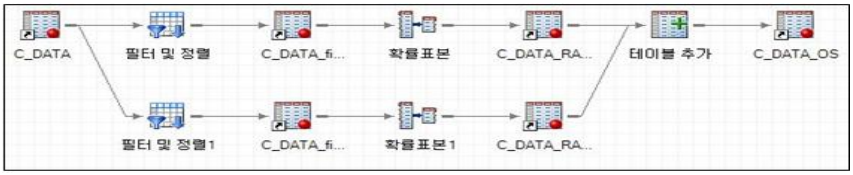

Figure 2.3 Oversampling procedure of EG 4.3

3. 모형구축 및 모형평가

3.1. 기본 모형 분석흐름도

본 연구에서 신용대출과 카드대출의 모형을 구축하는 과정에서 사용되어진 소프트웨어는 SAS사에서 제공하고 있는 Enterprise Miner workstation 7.1이다 (이하 EM). EM은 SAS사에서 제안하는 데이터 마이닝 분석절차인 SEMMA(샘플링, 데이터 탐색, 데이터 변환, 모델링, 모형평가)를 따라 다이어그램에 각 분석 노드들을 사용하여 분석흐름도를 작성할 수 있게 만든 데이터마이닝 분석도구이다.

본 연구에서는 신용대출과 카드대출의 모형을 구축하기 위하여 기본 모형만을 이용하여 분석을 시행한 기본 모형 분석흐름도와 분류 및 예측에 있어서 가장 고전적이지만 강력한 성능을 보이는 의사결정나무와 로지스틱 회귀모형에 앙상블 (ensemble) 기법인 배깅 (bagging)과 부스팅 (boosting)을 적용한 앙상블 모형 분석흐름도를 사용하여 분석을 진행하였다. 분석설계는 분석에 사용되어진 자료의 수(신용대출, 카드대출)와 모형을 구축하기 위해 원 데이터에서 분리하는 분석용과 검증용 표본의 추출 비율(본 연구에서는 6:4와 7:3), 그리고 마지막으로 분석기법(기본 모형 분석흐름도와 앙상블 모형 분석흐름도)에 따라서 Table 3.1과 같이 총 8개의 분석 다이어그램을 구성하였다.

Table 3.1 Design of analysis diagram

Data	Sample ratio	Model
Credit loan	6:4	Basic
Credit loan	6:4	Ensemble
Credit loan	7:3	Basic
Credit loan	7:3	Ensemble
Card loan	6:4	Basic
Card loan	6:4	Ensemble
Card loan	7:3	Basic
Card loan	7:3	Ensemble

분석 다이어그램의 유형은 크게 두 가지로 구분되며, 그 기준은 분석에 사용된 모형들의 종류이다. 즉, 분석 다이어그램은 기본 모형과 앙상블 모형으로 구분되며 Figure 3.1은 한 가지 예로서 분석설계에서 기본 모형만을 사용한 4 개의 다이어그램 중 하나이다.

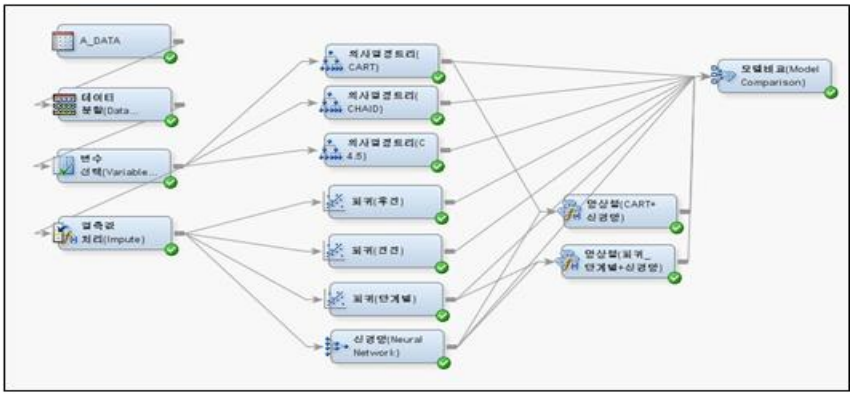

Figure 3.1 Example of analysis diagram for the basic model

Figure 3.1의 기본 모형 분석흐름도를 보면 분석노드들 사이사이마다 끝 모양이 화살표를 가진 직선

들이 연결되어있다. 이는 분석이 흘러가는 방향을 나타내며 전체적인 분석의 흐름은 좌측 상단에서 우측 상단으로 이동한다. 분석흐름도를 구성하고 있는 각각의 분석노드들에 대한 설명은 다음과 같다.

3.1.1. 데이터 분할

모형을 구축하기 위해서는 원 데이터에서 분석용 데이터 (training data)와 검증용 데이터 (validation data)로 나누어 주어야 한다. 여기서 분석용 데이터와 검증용 데이터의 크기 비율에 따라 모형의 분류 및 예측의 성능이 달라지기 때문에 본 연구에서는 분석용과 검증용의 데이터 셋을 정해주기 위하여 일반적으로 많이 사용하는 분할비율인 6:4와 7:3 등 두 가지 경우만을 고려하여 원 데이터에서 무작위 표본추출을 통해 데이터 셋을 구성하였다.

3.1.2. 변수선택

대용량 데이터에서 하나의 목표변수에 후보가 될만한 입력변수는 대단히 많이 존재하는 것이 일반적이며, 만약 입력변수를 모형구축에 포함시키게 되면 모형은 매우 복잡해지게 되고 모형자체가 전혀 무의미한 결과를 가져올 수 도 있다. 모형의 예측력이 중요하지만 간단하면 간단할수록 좋은 모형의 간명성도 또한 중요하다. EM에서는 입력변수 선택에 대한 옵션으로 결정계수와 카이제곱 통계량을 제공하고 있다. 본 연구에서는 결정계수를 이용하여 변수를 선택하였다.

3.1.3. 결측값 처리

앞서 설명하였듯이 모형구축에 있어서 데이터에 결측값이 존재하게 되면 해당 관찰치의 모든 값이 분석에서 제외되기 때문에 생성된 모형이 불안정해지게 된다. 따라서 결측값은 적절한 값으로 대체하는 것이 바람직하다. 본 연구에서는 입력변수가 범주형인 경우 결측값에 대한 보간을 최빈값을 사용하였고, 연속형인 경우에는 이상치가 많은 데이터 특성을 고려하여 중위수를 사용하였다. 또한 결측값을 하나의 범주로 보는 의사결정나무 모형의 특성을 고려하여 분석흐름도의 구성에 있어서 결측값 처리는 의사결정나무 모형을 제외한 모형들에게만 처리될 수 있도록 분석흐름도를 작성하였다.

3.1.4. 분석모형

본 연구에서 기본 모형 분석흐름도에 사용된 모형의 종류로는 크게 의사결정나무와 로지스틱 회귀모형, 신경망이 있다. 의사결정나무의 대표적 알고리즘에는 Breiman (1984)의 CART (classification and regression trees; gini), Hartigan (1975)의 CHAID (chi-squared automatic interaction detection), Quinlan (1993)의 C4.5 등의 알고리즘있으며, 최근 Park (2010), Cho와 Park (2011a, 2011b) 등은 의사결정나무 모형의 모형구축 시간단축 및 생성모형의 정확성 제고를 위한 하이브리드 데이터마이닝 방법론을 제안하였다. 본 연구에서는 의사결정나무의 분리규칙 옵션을 달리하여 CART(gini), CHAID, C4.5 등 세 가지의 모형으로 구성하였고, 로지스틱 회귀모형은 모형선택 방법에 따라 후진제거, 전진선택, 단계별선택 등 세 가지의 모형으로 구성하였다. 신경망 모형은 다층 퍼셉트론 (multi layer perceptron; MLP)을 사용하였고 은닉마디의 수는 3개이다. 마지막으로 CART와 신경망 모형이 결합한 모형, 로지스틱 회귀모형(단계적선택)과 신경망 모형이 결합한 모형 총 9개의 모형을 구성하였다. 여기서 모형의 결합에 사용되어지는 노드는 앙상블 노드이며 두 모형의 사후확률에 평균값을 이용한다.

3.1.5. 모형비교 방법

하나의 데이터를 분석할 때는 여러 통계모형들을 상정하여 분석하는 것이 바람직하다. 왜냐하면, 바람직한 통계분석은 구축된 모형에 따른 결과를 비교하여 최적의 모형을 선택해야 하기 때문이다. 최

적의 모형을 얻기 위해서는 여러 모형을 비교, 평가해야하고, 이를 통해 하나의 모형이 선택되면 선택된 모형이 다른 모형에 비해 우수하다는 것을 입증해야한다. 모형평가 노드는 분석흐름도에 사용되어진 모형들에 대한 ROC (receiver operating characteristic) 곡선과 반응률 및 예측력을 비교하여 모형선택 기준에 따라 최적의 모형을 선택하여 준다. 본 연구에서는 분석에 사용된 9개의 모형을 ROC 곡선을 이용하여 1차 평가하였고 오분류율 (misclassification rate), 반응률 (response rate), 향상도 (improvement degree), 누적 반응 검출률 (cumulative response detection rate)을 통해 2차 평가하였으며, 모형선택 기준으로는 오분류율을 사용하였다.

3.2. 앙상블 모형 분석흐름도

분석 다이어그램의 유형 중 두 번째 유형은 앙상블 모형 분석흐름도이다. Figure 3.2는 4개의 앙상블 모형 분석흐름도 중 한 개이다. 앙상블 모형 분석흐름도의 전반적인 절차는 기본 모형 분석흐름도에서 설명했던 것과 동일하다. 앙상블 모형 분석흐름도에서 사용되어진 모형은 분류규칙이 CART(gini)인 의사결정나무 모형과 로지스틱회귀 모형이다.

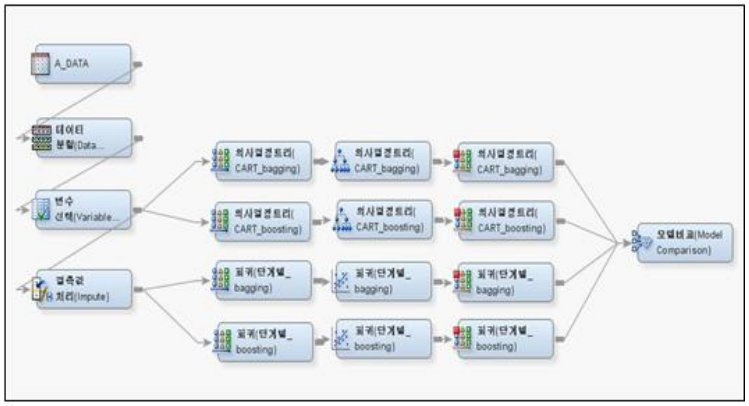

Figure 3.2 Example of analysis diagram for the ensemble model

3.2.1. 그룹시작

반복 구간의 시작 지점을 설정할 수 있으며, 본 연구에서는 이 노드를 이용하여 의사결정나무 모형과 로지스틱회귀 모형에 각각 배깅과 부스팅을 적용하였고 반복 횟수는 10회로 정하였다.

3.2.2. 그룹종료

반복 구간의 종료 지점을 설정하여 준다. 반복이 끝난 후의 결과를 다음 분석노드로 전달하여 준다.

3.3. 기본 모형 결과 비교

3.3.1. 신용대출

ROC 곡선은 구축 모형의 성능을 민감도와 특이도에 의해 판단하고자 하는 곡선이다. ROC 곡선을 그리기 위해서는 0과 1사이의 값을 갖는 분류 기준값에서의 민감도와 특이도를 계산하여, 수직축은 민감도, 수평축은 1-특이도로 하여 그릴 수 있다. ROC 곡선의 넓이를 나타내는 ROC 인덱스가 증가할수록 모형의 성능이 좋다고 판단한다. 본 연구에서는 모형의 평가에 있어서 앞서 설명하였듯이 1차적으로

ROC 곡선으로 모형의 성능을 판단하고, 2차적으로는 향상도, 반응율, 누적 반응 검출률을 이용하여 모형의 성능의 종합적인 관점에서 살펴보았으며, 마지막으로 모형 구축에서 가장 중요한 예측의 정확성을 볼 수 있는 오분류율을 이용하여 최종모형을 선별하였다.

Figure 3.3은 분석용과 검증용 데이터 셋의 비율이 각각 6:4와 7:3인 데이터에서 기본 모형 9가지를 이용하여 ROC 곡선을 그려본 것이다. 분석용과 검증용 데이터 셋의 비율 관점에서 ROC 곡선을 살펴보면 기본 모형 9가지의 ROC 곡선들이 크게 차이를 보이지 않는다. 두 번째로 일반화 관점에서 ROC 곡선을 살펴보면 분석용 데이터 셋에서 그려진 ROC 곡선의 모양과 검증용 데이터 셋에서 그려진 ROC 곡선의 모양이 많이 다르지 않는 것을 확인할 수 있다. 즉, ROC 곡선을 사용한 모형의 평가에 있어서 분석용과 검증용 데이터 셋의 비율이 모형의 성능에 크게 좌우하지 않는 것을 확인할 수 있다.

(6:4)

(7:3)

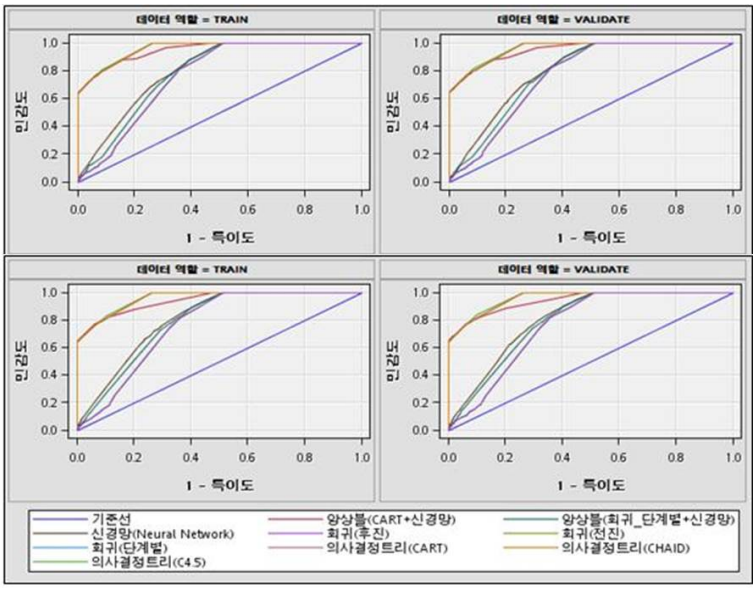

Figure 3.3 ROC curves (Basic model/Credit loan)

Table 3.2는 ROC 곡선의 넓이를 나타내는 ROC 인덱스를 각각의 기본 모형별로 정리해 놓은 것이다. 여기서 Tree, Tree2, Tree3은 순차적으로 CART(gini), CHAID, C4.5(entropy) 분류규칙을 사용한 의사결정나무 모형이다. 다음으로 Reg, Reg2, Reg3는 순차적으로 후진제거, 전진선택, 단계별선택 모형선택 옵션을 사용한 로지스틱회귀 모형이다. Neural은 다층퍼셉트론을 이용한 신경망 모형이며, Ensmbl는 CART(gini)와 신경망을 결합한 모형이며, Ensmbl2는 로지스틱 회귀(단계별선택)과 신경망을 결합한 모형이다. 샘플링 방법에 상관없이 C4.5 모형이 ROC 인덱스가 가장 높음을 보였다.

Table 3.2 ROC index (Basic model/Credit loan)

Sample ratio	Tree	Tree2	Tree3	Reg	Reg2	Reg3	Neural	Ensmbl	Ensmbl2
6:4	0.958	0.958	0.960	0.764	0.764	0.764	0.801	0.948	0.784
7:3	0.958	0.958	0.960	0.765	0.765	0.765	0.807	0.947	0.791

Table 3.3은 ROC 도표와는 다른 개념의 평가 도구인 이익 도표이다. 여기서 모형평가를 위해서 사용한 통계량은 향상도 (improvement degree), 반응률 (response rate), 누적 반응 검출률 (cumulative

response detection rate)이며, 데이터 셋을 10등분 하였을 때 상위 10%의 이익 도표 결과이다. 여기서 사용된 통계량에 대한 자세한 설명은 Kang과 Han (1999)을 참조하면 알 수 있다. Table 3.3을 자세히 살펴보면 ROC 도표와 달리 분석용과 검증용의 표본비율에 따라 결과의 차이를 보인다. 표본비율이 6:4일 경우 Tree3 모형이 향상도 (6.10301), 반응률 (85.86772), 누적 반응 검출률 (66.06232)에서 가장 좋은 성능을 보였으며, 표본비율이 7:3일 경우에는 Ensmbl 모형이 향상도 (6.15709), 반응률 (86.62815), 누적 반응 검출율 (66.32937)에서 가장 좋은 성능을 보였다.

Table 3.3 Profit table (Basic model/Credit loan)

Sample ratio	Model	Improvement degree	Response rate	Cumulative response detection rate
6:4	Tree	6.00937	84.55023	65.59413
	Tree2	6.00937	84.55023	65.59413
	Tree3	6.10301	85.86772	66.06232
	Reg	1.31053	18.43880	14.94932
	Reg2	1.31053	18.43880	14.94932
	Reg3	1.31053	18.43880	14.94932
	Neural	2.25318	31.70170	23.69240
	Ensmbl	6.01889	84.68409	65.64169
	Ensmbl2	1.40342	19.74574	18.44095
7:3	Tree	5.97807	84.10935	65.43437
	Tree2	5.97807	84.10935	65.43437
	Tree3	6.07120	85.41966	65.89996
	Reg	1.31466	18.49677	15.12685
	Reg2	1.31466	18.49677	15.12685
	Reg3	1.31466	18.49677	15.12685
	Neural	2.12399	29.88388	24.33792
	Ensmbl	6.15709	86.62815	66.32937
	Ensmbl2	1.97672	27.81171	21.53448

Table 3.4는 기본 모형 9개의 오분류율을 정리한 것이다. 모형의 평가에 있어서 아무리 안정적이고 효과적인 모형도 실제 문제에 적용했을 경우 빗나간 결과만을 양산한다면 아무런 의미가 없을 것이다. 따라서 구축된 모형이 얼마나 예측과 분류에서 뛰어난 성능을 보이는 지 판단하기 위하여 각 모형의 오분류율을 살펴보았고 본 연구에서는 이를 최종적인 모형 선택기준으로 사용하였다. Table 3.4에서는 앞서 살펴 보았던 ROC 도표, 이익 도표와 달리 Tree 모형이 표본비율 6:4와 7:3 기준으로 하여 각각의 오분류율이 0.05186, 0.05186으로 가장 낮음을 보였다.

Table 3.4 Misclassification rate (Basic model/Credit loan)

Sample ratio	Tree	Tree2	Tree3	Reg	Reg2	Reg3	Neural	Ensmbl	Ensmbl2
6:4	0.05186	0.05200	0.05206	0.14720	0.14720	0.14720	0.14071	0.05226	0.14104
7:3	0.05185	0.05185	0.05202	0.14746	0.14746	0.14746	0.14066	0.05206	0.14129

3.3.2. 카드대출

Figure 3.4는 분석용과 검증용 표본비율에 따라 그려진 기본 모형들에 대한 ROC 곡선이며, 카드대출에서도 앞의 신용대출과 마찬가지로 분석용과 검증용의 샘플링비율 관점에서 ROC 곡선의 성능에 차이를 큰 차이를 보이지 않는다. 또한 모형의 일반화의 관점에서도 ROC 곡선 성능에 차이를 보이지 않음을 알 수 있었다.

(6:4)

(7:3)

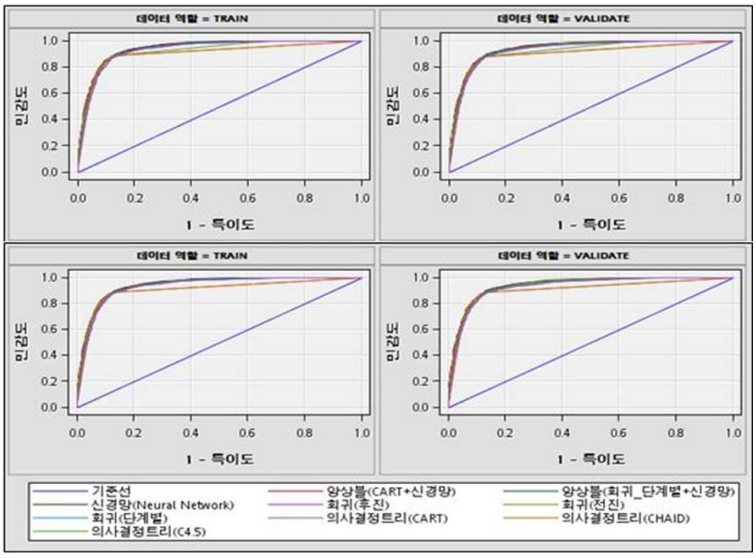

Figure 3.4 ROC curves (Basic model/Card loan)

Table 3.5는 ROC 인덱스를 정리해 놓은 것이다. 표본비율에 관계없이 Ensmbl 모형의 ROC 인덱스가 0.941로 가장 높음을 보였다.

Table 3.5 ROC index (Basic model/Card loan)

Sample ratio	Tree	Tree2	Tree3	Reg	Reg2	Reg3	Neural	Ensmbl	Ensmbl2
6:4	0.900	0.901	0.924	0.929	0.935	0.929	0.940	0.941	0.937
7:3	0.901	0.901	0.937	0.928	0.935	0.928	0.940	0.941	0.937

Table 3.6은 향상도, 반응률, 누적 반응 검출률을 이용하여 각 9개의 모형을 평가한 이익 도표이다. 표본비율이 6:4인 경우 Ensmbl 모형이 향상도 (4.0326), 반응률 (80.65), 누적 반응 검출률 (42.375)에서 가장 좋은 성능을 보였고, 표본비율이 7:3인 경우에도 마찬가지로 Ensmbl 모형이 향상도 (4.0318), 반응률 (80.63333), 누적 반응 검출률 (42.424)에서 가장 좋은 성능을 보임을 알 수 있었다.

Table 3.6 Profit table (Basic model/Card loan)

Sample ratio	Model	Improvement degree	Response rate	Cumulative response detection rate
6:4	Tree	3.82459	76.48984	38.72134
	Tree2	3.82459	76.48984	39.12717
	Tree3	3.99445	79.88700	42.01790
	Reg	3.90135	78.02500	39.75625
	Reg2	3.83010	76.60000	41.13750
	Reg3	3.90135	78.02500	39.75625
	Neural	3.99260	79.85000	42.18750
	Ensmbl	4.03260	80.65000	42.37500
	Ensmbl2	3.91426	78.28333	41.29583
7:3	Tree	3.81776	76.35275	39.08319
	Tree2	3.81776	76.35275	38.90299
	Tree3	3.84091	76.81564	40.65959
	Reg	3.93096	78.61667	39.76667
	Reg2	3.85346	77.06667	41.20000
	Reg3	3.93096	78.61667	39.76667
	Neural	4.01513	80.30000	42.34167
	Ensmbl	4.03180	80.63333	42.52500
	Ensmbl2	3.93680	78.73333	41.23333

오분류율을 나타낸 Table 3.7을 보면, Ensmbl 모형에서 표본비율이 6:4인 경우 오분류율이 0.10122, 7:3인 경우 오분류율이 0.10083으로 모두 가장 낮은 오분류율을 보였다.

Table 3.7 Misclassification rate (Basic model/Card loan)

표본비율	Tree	Tree2	Tree3	Reg	Reg2	Reg3	Neural	Ensmbl	Ensmbl2
6:4	0.10235	0.10245	0.10287	0.11052	0.10957	0.11052	0.10430	0.10122	0.10365
7:3	0.10190	0.10206	0.10233	0.11090	0.10856	0.11090	0.10446	0.10083	0.10300

3.4. 앙상블 모형 결과 비교

3.4.1. 신용대출

앙상블 모형에서는 CART(gini) 분류 규칙을 사용한 의사결정나무 모형과 로지스틱회귀(단계적선택) 모형에 각각 배깅과 부스팅을 적용한 모형을 구축하였다. Figure 3.5의 ROC 곡선을 살펴보면 표본비율과 일반화 기준에서 ROC 곡선에 큰 차이를 보이지 않으며, 의사결정나무 모형이 로지스틱 회귀모형보다 성능이 우수한 것을 알 수 있다.

(6:4)

(7:3)

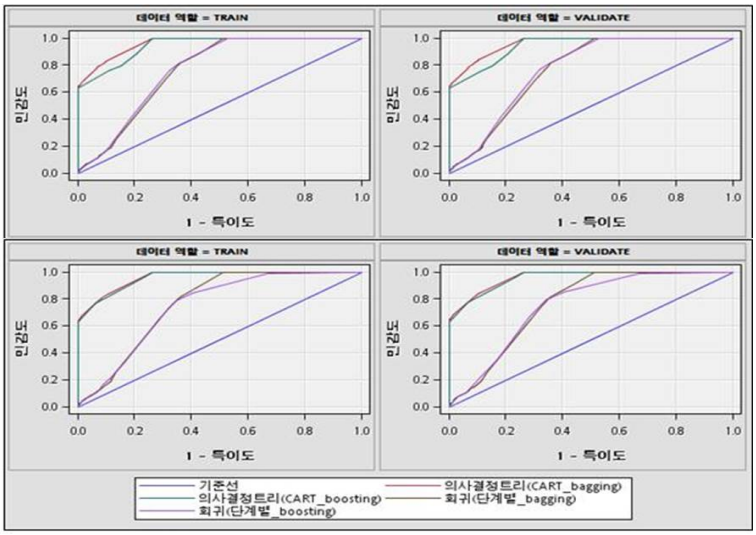

Figure 3.5 ROC curves (Ensemble model/Credit loan)

Table 3.8의 ROC 인덱스를 살펴보면 표본비율이 6:4인 경우 의사결정나무 모형에서 ROC 인덱스가 배깅 (0.962)이 부스팅 (0.943) 보다 높았고, 로지스틱 회귀모형에서는 부스팅 (0.772)이 배깅 (0.763)보다 높음을 보였다. 표본비율이 7:3인 경우 의사결정나무 모형의 ROC 인덱스가 배깅 (0.961)이 부스팅 (0.955) 보다 높았으며, 로지스틱 회귀모형에서는 배깅 (0.765)이 부스팅 (0.755)보다 높음을 보였다. 따라서 ROC 도표를 이용한 모형 평가에서 의사결정나무 모형이 로지스틱 회귀모형 보다 더 좋은 성능을 보였고 배깅과 부스팅 모형에서 전반적으로 배깅이 더 좋은 성능을 보여주는 것을 알 수 있다.

Table 3.9의 이익 도표를 살펴보면, 표본비율이 6:4인 경우 배깅을 사용한 의사결정나무 모형이 향상도 (6.12131), 반응률 (86.12515), 누적 반응검출률 (66.15380)이 가장 높고, 표본비율이 7:3인 경우도 마찬가지로 배깅을 사용한 의사결정나무 모형이 향상도 (6.12235), 반응률 (86.13935), 누적 반응 검출률 (66.15569)이 가장 높았다. 즉, 표본비율에 관계없이 의사결정나무 모형이 로지스틱 회귀모형보다 좋은 성능을 보였으며, 부스팅보다는 배깅이 더 좋은 성능을 보임을 알 수 있었다.

Table 3.8 ROC index (Ensemble model/Credit loan)

Sample ratio	Decision Tree		Regression	
	Bagging	Boosting	Bagging	Boosting
6:4	0.962	0.943	0.763	0.772
7:3	0.961	0.955	0.765	0.755

Table 3.9 Profit table (Ensemble model/Credit loan)

Sample ratio	Model		Improvement degree	Response rate	Cumulative response detection rate
6:4	Decision Tree	Bagging	6.12131	86.12515	66.15380
		Boosting	5.69812	80.17103	64.03790
	Regression	Bagging	1.32870	18.69449	15.04736
		Boosting	1.37825	19.39156	15.18031
7:3	Decision Tree	Bagging	6.12235	86.13935	66.15569
		Boosting	5.89001	82.87046	64.99415
	Regression	Bagging	1.32318	18.61670	15.11952
		Boosting	1.57956	22.22381	16.64956

Table 3.10의 오분류율을 살펴보면, 표본비율이 6:4인 경우 배깅을 사용한 의사결정나무 모형의 오분류율이 0.05186으로 가장 낮았으며, 표본비율이 7:3 경우도 배깅을 사용한 의사결정나무 모형의 오분류율이 0.05189로 가장 낮았다.

Table 3.10 Misclassification rate (Ensemble model/Credit loan)

Sample ratio	Decision Tree		Regression	
	Bagging	Boosting	Bagging	Boosting
6:4	0.05186	0.19489	0.14707	0.24314
7:3	0.05189	0.05214	0.14748	0.14309

3.4.2. 카드대출

Figure 3.6의 ROC 곡선을 살펴보면, 카드대출에서 앙상블 모형은 일반화 관점에서 ROC 곡선의 차이는 없었지만, 분석용과 검증용의 표본비율 관점에서는 차이를 보였다. 표본비율 7:3에서 부스팅을 적용한 의사결정나무 모형의 ROC 곡선이 표본비율 6:4인 경우보다 면적이 더 작음을 보였다.

(6:4)

(7:3)

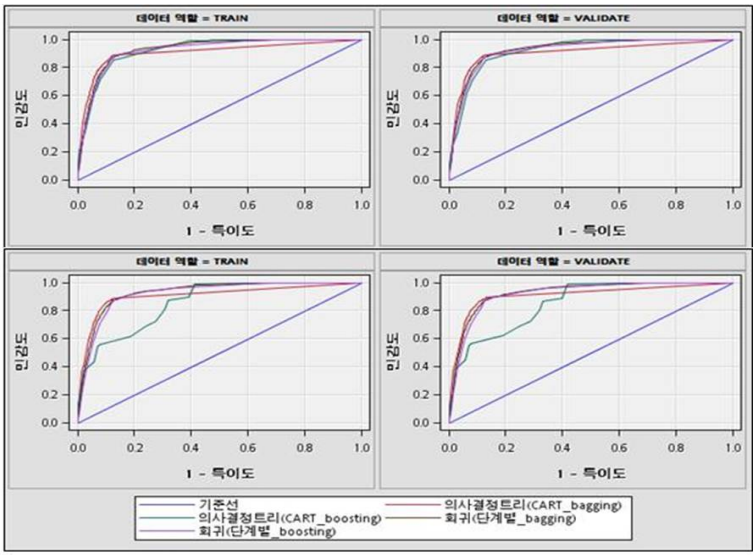

Figure 3.6 ROC curves (Ensemble model/Card loan)

Table 3.11을 보면 표본비율이 6:4인 경우 부스팅을 사용한 의사결정나무의 ROC 인덱스가 0.926으로 표본비율이 7:3인 경우 부스팅을 사용한 의사결정나무의 ROC 인덱스 0.855보다 크며, 다른 모형들의 차이보다 더 큰 차이를 보임을 알 수 있었다. 또한 ROC 도표를 이용한 모형평가에서는 전반적으로 배깅이 부스팅보다 모형 성능이 우수하였고 표본비율 6:4, 7:3에서 부스팅을 적용한 로지스틱 회귀분석의 ROC 인덱스가 각각 0.929, 0.928로 가장 높았다.

Table 3.12의 이익 도표를 살펴보면, 표본비율이 6:4인 경우 배깅을 사용한 의사결정나무 모형의 향상도와 반응률, 누적 반응 검출률이 각각 4.02868, 80.57159, 41.74665로 가장 높았으며, 표본 비율이 7:3인 경우 배깅을 적용한 의사결정나무 모형의 향상도와 반응률, 누적 반응 검출률이 각각 3.93371, 78.67161, 41.79609로 가장 높았다. 이익 도표의 모형평가에서 전반적으로 배깅이 부스팅보다 우수하였으며 표본비율이 6:4인 경우 모형이 더 좋은 성능을 보였다.

Table 3.11 ROC index (Ensemble model/Card loan)

	Decision Tree		Regression	
Sample ratio	Bagging	Boosting	Bagging	Boosting
6:4	0.909	0.926	0.929	0.924
7:3	0.924	0.855	0.928	0.921

Table 3.12 Profit table (Ensemble model/Card loan)

Sample ratio	Model		Improvement degree	Response rate	Cumulative response detection rate
6:4	Decision Tree	Bagging	4.02868	80.57159	41.74665
		Boosting	2.96887	59.37597	35.94923
	Regression	Bagging	3.89920	77.98200	39.78750
		Boosting	3.80760	76.15000	39.22500
7:3	Decision Tree	Bagging	3.93371	78.67161	41.79609
		Boosting	3.76259	75.24936	38.87727
	Regression	Bagging	3.91680	78.33333	39.73333
		Boosting	3.74346	74.86667	38.36667

Table 3.13의 오분류율을 살펴보면, 표본비율이 6:4인 경우 배깅을 적용한 의사결정나무 모형의 오분류율이 0.10252로 다른 모형보다 가장 낮았으며, 표본비율이 7:3인 경우 배깅을 적용한 의사결정나무 모형의 오분류율이 0.10156으로 가장 낮았다. 오분류율 표에서 표본비율이 6:4인 경우 보다 7:3인 경우가 더 좋은 성능을 보였다.

Table 3.13 Misclassification rate (Ensemble model/Card loan)

	Decision Tree		Regression	
표본비율	Bagging	Boosting	Bagging	Boosting
6:4	0.10252	0.33182	0.11187	0.12852
7:3	0.10156	0.65011	0.11180	0.13373

4. 결론

본 연구에서는 국내 여신기관의 신용대출과 카드대출 관련 실제 자료에 대하여 SAS EM에서 제안하는 데이터마이닝 분석절차인 SEMMA를 따라 데이터 탐색에서부터 구축모형의 비교 및 평가까지 데이터마이닝의 전반적인 과정에 대하여 고찰하였다.

본 연구의 첫 번째 주안점은 데이터 사전처리로서, 목표변수가 사례집단과 대조집단의 비율이 1:99인 카드대출 데이터에 오버샘플링을 적용하여 목표변수의 비율을 1:4로 재구성하였으며, 재구성된 데이터를 사용하여 다양한 모형을 구축하고 성능을 비교한 결과 오버샘플링된 데이터가 모형구축에 적합함을 보였다.

두 번째 주안점은 모형구축에 있어서 다양한 모델링 기법을 사용하고, 여러 가지 측면에서 모형을 비교, 평가하여 최적의 모형을 도출하는 과정을 설명하는 것이다. 본 연구에서는 기본 모형으로 의사결정나무 모형, 로지스틱회귀 모형, 신경망 모형을 사용하였고, 신경망 모형과 의사결정나무 모형 (CART), 신경망 모형과 로지스틱회귀 모형 (단계별선택)을 결합하고 배깅과 부스팅을 적용한 앙상블 모형을 4개 구성하였다. 모형평가 결과, 반응률 관점으로 볼 때 신용대출 자료에서는 표본비율이 7:3일 때 더 잘 적합하였으며, 카드대출 자료에서는 표본비율이 6:4일 때 더 잘 적합하였다. 그리고 반응률은 기본 모형이 앙상블 모형보다 더 높았으며 앙상블 모형에서는 배깅이 부스팅보다 더 높은 반응률을 보였다.

본 연구에서 사용한 모형 선택기준인 오분류율에서 모형을 평가하면 전반적으로 7:3의 표본 분할 기준이 가장 적합하였으며, 분류규칙으로는 CART 모형이 본 데이터에 가장 작 적합하는 것을 알 수 있었다. 또한 배깅이 부스팅보다 더 낮은 오분류율을 보였지만 기본 모형 보다는 높은 오분류율을 보였다.

이처럼 본 연구에서는 표본비율을 달리하고 여러 가지 모델링 기법을 사용하여 2개의 데이터에 모형들을 구축하고 비교 및 평가를 하는 일련의 과정을 정립하였다. 또한 생성된 모형들이 한 가지의 평가방법으로 평가되기 보다는 여러 가지 평가방법을 종합하여 판단하는 것이 모형 평가 방법에 있어서 더 안정적임을 보였고, 여기에 부가적으로 샘플링 비율에 따라 모형의 성능이 크지는 않지만 조금씩 차이를 보이는 것과 오분류율을 줄여주는 앙상블 모형이 기본 모형보다 항상 좋지 않다는 것을 알 수 있었다.

References

Berry, M. and Linoff, G. (1997). *Data mining techniques: For marketing, sales and customer support*, Wiley, New York.

Berry, M. and Linoff, G. (2011). *Data mining techniques: For marketing, sales and customer relationship management*, Wiley, New York.

Breiman, L. (1984). *Algorithm CART*, California Wadsworth International Group, Belmont, CA.

Cho, K. H. and Park, H. C. (2011a). A study on decision tree creation using intervening variable. *Journal of the Korean Data & Information Science Society*, **22**, 671-678.

Cho, K. H. and Park, H. C. (2011b). A study on removal of unnecessary input variables using multiple external association rule. *Journal of the Korean Data & Information Science Society*, **22**, 877-884.

Chung, H., Kang, C. and Kim, K. C. (2008). A study on the effect of oversampling for unbalanced data. *Journal of the Korean Data Analysis Society*, **10**, 2089-2098.

Hartigan, J. A. (1975). *Algorithm CHAID*, John Wiley and Sons, New York.

Kang, H and Han, S. (1999). *Data mining methodology and application*, Free-Academy, Seoul.

Park, H. C. (2010). Association rule ranking function by decreased lift influence. *Journal of the Korean Data & Information Science Society*, **21**, 397-405.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*, San Mateo, Morgan Kaufmann, CA.

Journal of the Korean Data & Information Science Society
2013, **24**(3), 411–425

http://dx.doi.org/10.7465/jkdi.2013.24.3.411
한국데이터정보과학회지

A study on the analysis of customer loan for the credit finance company using classification model

Tae-Hyung Kim[1] · Yeong-Hwa Kim[2]

[1]SAS Korea
[2]Department of Applied Statistics, Chung-Ang University

Received 12 March 2013, revised 17 April 2013, accepted 22 April 2013

Abstract

The importance and necessity of the credit loan are increasing over time. Also, it is a natural consequence that the increase of the risk for borrower increases the risk of non-performing loan. Thus, we need to predict accurately in order to prevent the loss of a credit loan company. Our final goal is to build reliable and accurate prediction model, so we proceed the following steps: At first, we can get an appropriate sample by using several resampling methods. Second, we can consider variety models and tools to fit our resampling data. Finally, in order to find the best model for our real data, various models were compared and assessed.

Keywords: Big data, data mining, decision tree, loan, oversampling, risk management.

[1] Consultant, SAS Korea, Daechi-Dong 889, Gangnam-Gu, Seoul 135-839, Korea
[2] Corresponding author: Professor, Department of Applied Statistics, Chung-Ang University, Seoul 156-756, Korea. E-mail: gogators@cau.ac.kr