Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 1, pp. 1 ~ 10

Churn Analysis for the First Successful Candidates in the Entrance Examination for K University

Kyu Il Kim[1] · Seung Han Kim[2] · Eun Young Kim[3]
Hyun Kim[4] · Jae Wan Yang[5] · Jang Sik Cho[6]

Abstract

In this paper, we focus on churn analysis for the first successful candidates in the entrance examination on 2006 year using Clementine, data mining tool. The goal of this study is to apply decision tree including C5.0 and CART algorithms, neural network and logistic regression techniques to predict a successful candidate churn. And we analyze the churning and nochurning successful candidates and why the successful candidates churn and which successful candidates are most likely to churn in the future using data from entrance examination data of K university on 2006 year.

Keywords : 데이터 마이닝, 로지스틱 회귀모형, 신경망 모형, 예측모형, 의사결정나무 모형

1. 서론

자원의 발달로 인하여 무한 경쟁사회로 접어든 이 사회의 환경을 살펴보면 시장경제의 심화와 빠른 변화를 확인 할 수 있다. 그리고 경영인들의 관심은 '어떤 일이 일

1) Student, Department of Informational Statistics, Kyungsung University, Busan, 608-736, Korea. E-Mail : kkiyr@hanmail.net
2) Student, Department of Informational Statistics, Kyungsung University, Busan, 608-736, Korea. E-Mail : kuduckno1@naver.com
3) Student, Department of Informational Statistics, Kyungsung University, Busan, 608-736, Korea. E-Mail : click170@nate.com
4) Student, Department of Informational Statistics, Kyungsung University, Busan, 608-736, Korea. E-Mail : cbsangelqq@nate.com
5) Student, Department of Informational Statistics, Kyungsung University, Busan, 608-736, Korea. E-Mail : jy2lemona@nate.com
6) (Corresponding Author) Associate Professor, Department of Informational Statistics, Kyungsung University, Busan, 608-736, Korea. E-Mail : jscho@ks.ac.kr

어났는가?'에서 '앞으로 어떤 일들이 일어날 것인가?'로 점점 바뀌고 있으며, 이러한 환경에서 예측의 중요성은 점점 더 강조된다. 또한 현대 사회는 기술의 발달로 인해 많은 양의 데이터들을 축적할 수 있으며, 이렇게 축적된 많은 양의 데이터로 부터 앞으로 어떤 일들이 일어날 것인지에 대한 예측모형을 만든다면 그 데이터들은 단지 기록의 역할을 넘어서서 그 방향을 제시해 줄 수 있는 중요한 길잡이 역할도 가질 수 있게 된다. 따라서 대용량의 데이터에 내재되어 있는 관계, 패턴 및 규칙들을 찾아내고 이를 모형화하여 유용한 정보로 변환하는 일련의 과정으로써 데이터마이닝 기법의 사용은 많은 분야에서 적용이 가능하다. 데이터마이닝에 대한 자세한 내용은 최종후 등(1998), 허명회 등(2003), 허준 등(2003) 및 허준 등(2001)을 참고하기 바란다.

한편, 앞으로의 대학입시는 격감하는 인구감소에 따른 대량의 미달사태 뿐만 아니라 복수지원이 허용되는 현행 입시제도 때문에 한 학생이 여러 대학에 합격하는 경우가 발생하게 된다. 이런 경우, 수험생들은 합격한 대학들 중에서 원하는 한 대학을 선택해서 등록을 하기 때문에 많은 대학들의 최초합격자들은 미등록, 등록포기, 환불 등의 이유에 의해 타 대학으로 유출되는 이탈자 문제로 어려움을 겪고 있다.

따라서 최초 합격자 발표에서 합격한 학생이라 할지라도 이 학생이 합격한 그 대학에 등록을 할 학생인지 아닌지는 기존의 입시분석 보고서를 통해서 알 수 있는 근거가 없는 것이 현실이며, 또한 대학의 입시관리 부서에서는 이탈자로 생긴 결손 인원만큼 추가 합격자로 충당해야 하므로 많은 시간과 비용 및 행정적인 로드가 걸릴 뿐만 아니라 최초합격자의 우수 학생들이 타 대학으로 유출되기 때문에 최초합격자를 대신해서 입학하게 되는 추가 합격자들은 최초합격자들 보다 학업능력이 떨어지는 것이 현재 대부분 대학들이 겪고 있는 어려움이라고 할 수 있다.

따라서 본 연구에서는 데이터마이닝 기법을 이용하여 2006년 국내 K 대학교에 지원한 수험생들 중 학업능력이 뛰어난 최초합격자들이 타 대학으로 이탈한 학생들의 성향을 분석하고 이를 통해 이탈자들에 대한 예측모형을 개발하여 입시관리부서에서 이탈자들에 대한 효과적인 전략을 수립하는데 기초자료를 제공하고자 한다. 이를 위해 의사결정나무 모형, 신경망 모형, 로지스틱 회귀 모형 등으로 이탈자에 대한 예측모형을 구축하고, 이익도표를 이용한 각 모형의 성능을 비교·평가를 통해 보다 나은 예측모형을 제시하고자 한다.

2. 자료설명

본 연구에서 사용되는 분석 자료는 K 대학의 2006학년도 입시자료이며, '단과대학', '성별', '수능점수'등 변수의 수가 38개로서, 관찰 값의 수는 21,022개이다. 여기서 정시 '가'군과 '나'군은 전형방법이 동일하며, 최초 합격자의 수는 925명이며, 본 연구에서는 정시 '가'와 정시 '나'의 최초합격자들의 이탈자에 대한 예측모형을 구축하고자 한다. 여기서 최초 합격자가 최종 등록을 하는 경우를 '등록'이라고 하고, 최초에는 합격을 했지만 환불, 등록포기, 미등록 등의 이유로 최종적으로 K 대학에 등록하지 않은 경우를 '이탈'이라고 정의를 한다면, 최초합격자들 중 최종등록여부에 의해서 '이탈여부(이탈, 등록)'가 결정이 된다.

한편 38개의 변수들 중 중복 의미를 가지는 변수들과 예측모형 구축에 적합하지 않은 변수들도 다수 포함되어 있어서 분석목적에 부합하도록 전체 데이터에 대한 정제작업을 실시하였다. 또한 많은 설명변수들로 예측모형을 구축하게 되면 모형이 복잡

해질 뿐만 아니라, 구축된 모형을 해석하기 어렵다는 단점 때문에 목표변수와 관련성이 높은 설명변수를 선택할 필요가 있다. 따라서 목표변수에 통계적으로 유의한 영향을 미치는 설명변수만을 선택하였으며, 선택되어진 변수는 다음 <표 1>과 같다.

<표 1> 분석 데이터 셋 내의 변수

변수 이름	변수형태	변수 값
이탈여부(목표변수)	명목형	등록(0), 이탈(1)
모집전형	명목형	국가유공자(1), 내신일반(2), …, 학생임원・개근 자(16)
성별	명목형	남자(1), 여자(2)
재수여부	명목형	재수(1), 재수안함(2)
대학 명	명목형	공대(1), 멀대(2),…, 이대(9)
학과계열	명목형	예체능계(1), 인문계(2), 자연계(3)
탐구영역계열	명목형	과학탐구(1), 사회탐구(2), 직업탐구(3)
지역	명목형	부산(1), 경남(2), 기타(3)
내신성적평균	연속형	0~100
수능 백분위 점수	연속형	0~100

위의 <표 1>에서 이탈 여부에 대한 정확한 예측은 입시관리의 측면에서 매우 중요하며, 따라서 '이탈여부'를 목표변수로 하고 그 외의 모든 변수들을 설명변수로 하여 예측모형을 구축하고자 한다.

3. 모형구축

K 대학에서 최초에 합격한 수험생들 중 타대학으로 이탈한 학생들에 대한 예측모형을 구축하는 절차는 다음과 같다. 먼저 의사결정나무 분석, 로지스틱 회귀분석 및 신경망 분석을 이용하여 예측모형을 구축하였으며, 구축된 모형에 대해 이익도표를 이용하여 보다 나은 모형을 선택한 후, 선택된 모형에 대한 리프트(Lift)를 계산하여 가장 이탈가능성이 높은 집단을 선택하여 선택된 집단의 성향을 파악한다.

3.1 C5.0 알고리즘

의사결정나무분석의 알고리즘으로는 Kass(1980), Breiman et al.,(1984) 및 Quinlan (1993)에 의해 다양한 알고리즘이 제안되었다. 그 중에서 C5.0은 엔트로피 지수(entropy index)를 가장 감소시켜주는 설명변수와 그 변수의 최적분리를 자식마디로 분리하는 알고리즘으로서 엔트로피 지수는 다음과 같이 정의된다.

$$entropy = -\sum_{i=1}^{C} p_i \log(p_i), \quad (1)$$

여기서 C는 목표변수의 범주 수를 의미하며, p_i는 i번째 목표범주의 모비율을 의미한다. C5.0 알고리즘을 이용해서 의사결정나무 분석을 수행한 결과 의미있는 규칙집합은 다음 <표 2>와 같다.

<표 2> C5.0에 의한 모형구축 결과

집합 / 규칙	규칙집합
규칙 I	(학과계열 = 예체능계) & (대학명 = 멀대) & (탐구영역 = 사회탐구) & (지역 = 부산) & (내신성적 > 34.037) & (성별 = 여자) ⇒ 이탈(11, 0.909)
규칙 II	(학과계열 = 자연계) & (내신성적 ≤ 10.345) & (성별 = 남자) & (고교지역 = 부산) ⇒ 이탈(11, 0.818)
규칙 III	(학과계열 = 예체능계) & (대학명 = 예대) ⇒ 이탈(186, 0.672)

위의 결과에서 수험생들의 이탈여부에 영향을 미치는 변수를 살펴보면, 학과계열이 가장 중요한 변수이며, 학과계열에 따라서 대학명과 내신성적이 그 다음으로 중요하고 지역, 성별 등의 순으로 나타났다. 위의 결과에서 끝 노드(terminal node)의 괄호 안에 있는 두 개의 숫자는 해당 노드 내에서 관찰치의 빈도와 이탈비율을 의미한다.

이상의 <표 2>에 의한 분석결과를 근거로 이탈가능성이 높은 집단을 선별해 보면, 학과계열이 예체능계이면서 대학명이 멀대이고, 탐구영역이 사회탐구, 지역이 부산, 내신성적이 34.037 보다 높으면서 여자들임을 알 수 있는데, 이들 집단의 이탈율은 90.9%임을 알 수 있다. 또한 학과계열이 자연계이면서 내신성적이 10.345 이하이면서 성별이 남자이고, 부산지역에 거주하는 경우, 이 집단의 이탈율이 81.8%로 나타났고, 학과계열이 예체능이면서 대학명이 예대인 경우, 이들 집단의 이탈율이 67.2%임을 알 수 있다.

위의 모형에 대한 오분류율의 결과를 정리한 것이 아래 <표 3>과 같다.

<표 3> C5.0 알고리즘에 의한 오분류표

예측 \ 실제		실제범주		계
		등록	이탈	
예측범주	등록	0.43	0.12	0.55
	이탈	0.20	0.25	0.45
계		0.63	0.37	1.00

위의 결과를 보면 구축된 모형은 실제 이탈을 이탈로, 등록을 등록으로 제대로 예측한 경우는 전체 관찰치 중 68.0%로 나타났다.

3.2 CART 알고리즘

CART 알고리즘은 불순도(impurity) 또는 다양도(diversity)를 측정하는 지니지수(Gini index)를 사용하여 지니지수를 가장 감소시켜주는 설명변수와 그 변수의 최적분리를 자식마디로 분리한다. 기호 $P(i)P(j)$를 임의의 한 개체가 i 번째 범주로부터 추출되었고, 그 개체를 목표변수의 j 번째 범주에 속한다고 오분류할 확률이라고 한다면 지니지수는 다음과 같이 정의된다.

$$G = \sum_{j=1}^{C} \sum_{i \neq j} P(i)P(j) = 1 - \sum_{j=1}^{C} (n_j/n)^2, \quad (2)$$

여기서 n은 그 마디에 포함되어 있는 관찰치 수를 의미하고 n_i는 i번째 범주에 속하는 관찰치 수를 의미한다.

CART 알고리즘을 이용해 의사결정나무 분석을 수행한 결과 의미있는 규칙집합은 다음 <표 4>와 같다.

<표 4> CART 알고리즘에 의한 모형구축 결과

집합 / 규칙	규칙집합
규칙 I	(대학명 = 이대, 신대, 상대, 법대, 멀대) & (내신성적 < 6.138) & (탐구영역 = 직업탐구, 과학탐구) ⇒ 이탈(10, 1.0)
규칙 II	(대학명 = 이대, 상대, 문대, 공대) & (10.005 ≤ 내신성적 < 10.675) & (탐구영역 = 사회탐구) ⇒ 이탈(18, 0.722)
규칙 III	(대학명 = 이대, 신대, 상대, 법대, 멀대) & (6.138 < 내신성적 < 14.852) & (탐구영역 = 직업탐구, 과학탐구) ⇒ 이탈(22, 0.682)
규칙 IV	(대학명 = 예대) ⇒ 이탈(186, 0.672)

위의 결과에 따르면 수험생들의 이탈여부에 중요한 영향을 미치는 변수를 살펴보면, 대학 명, 내신성적, 탐구영역의 순으로 나타났으며, C5.0을 이용한 경우와 비교해 보면 다소간의 차이가 있음을 알 수 있다.

이상의 <표 4>에 의한 분석결과를 기초로 이탈가능성이 높은 집단을 선별해 보면, 이대, 신대, 상대, 법대, 멀대를 지원하면서 내신성적이 6.138보다 낮고 직업탐구 및 과학탐구를 선택한 경우를 들 수 있는데, 이들 집단의 이탈율은 100%임을 알 수 있다. 또한 이대, 상대, 문대, 공대를 지원하면서 내신성적이 10.005와 10.675 사이에 있으면서 사회탐구를 선택한 학생의 경우, 이들 집단의 이탈율은 72.2%로 나타났고, 또한 이대, 신대, 상대, 법대, 멀대를 지원하면서 내신성적이 6.138과 14.852 사이에 있고 직업탐구와 과학탐구를 선택한 경우, 이들 집단의 이탈율은 68.2%, 예술대학의 경우 이탈율이 67.2%로 나타났다.

위에서 적용한 모형에 대한 오분류율의 결과를 정리한 것이 아래 <표 5>와 같다.

<표 5> CART 알고리즘에 의한 오분류표

실제 / 예측		실제범주		계
		등록	이탈	
예측범주	등록	0.44	0.11	0.55
	이탈	0.23	0.22	0.45
계		0.67	0.33	1.00

위의 결과를 보면 구축된 모형은 실제 이탈을 이탈로, 등록을 등록으로 제대로 예측한 경우는 전체 관찰치 중 66.0%로 나타났다.

3.3 로지스틱 회귀모형

로지스틱 회귀모형은 목표변수가 이분형일 때 선형회귀모형의 단점을 극복하기 위해 사후확률에 대한 로짓변환(logit transformation)을 고려하여 분석하는 것으로서, 자세한 내용은 김순귀 등(2003)을 참고하기 바란다. 즉,

$$\ln\frac{p(y=1|x_1,\cdots,x_p)}{1-p(y=1|x_1,\cdots,x_p)}=\alpha+\beta_1x_1+\cdots+\beta_px_p \qquad (3)$$

와 같이 모형화 하여, 모형식의 좌변과 우변이 모두 실수 값을 가지도록 하는 것이다. 여기에서 α와 $\beta_1,\cdots,\beta_p$는 모수(parameter)들로서 추정되어야 할 회귀계수(regression coefficient)들이고, y는 목표변수의 관측 값이며, $x_1,\cdots,x_p$는 설명변수들이다.

로지스틱 회귀모형을 수행한 결과 통계적으로 유의한 설명변수만을 결과로 제시한 것이 다음 <표 6>과 같다.

<표 6> 로지스틱 회귀모형에 의한 모형구축 결과

설명변수	β 추정 값	유의확률	Exp(β)
내신 성적	0.018	0.002	1.018
[성별=남자]	0.449	0.004	1.566
[대학 명=예술대학]	-1.350	0.005	0.259
[지역=경남]	-0.476	0.019	0.621

위의 결과를 보면 수험생들의 이탈여부에 영향을 미치는 중요한 변수로는 내신성적, 성별, 대학 명, 지역으로서, 특히 성별이 남자인 경우, 대학명이 예대인 경우, 그리고 지역이 경남인 경우, 등록할 확률에 대한 이탈할 확률의 상대비가 각각 1.566배, 0.259배, 그리고 0.621배 증가함을 알 수 있다.

위에서 적용한 모형에 대한 오분류율의 결과를 정리한 것이 아래 <표 7>과 같다.

<표 7> 로지스틱 회귀모형에 의한 오분류표

실제 / 예측		실제범주		계
		등록	이탈	
예측범주	등록	0.45	0.10	0.55
	이탈	0.25	0.20	0.45
계		0.70	0.30	1.00

위의 결과를 보면 구축된 모형은 실제 이탈을 이탈로, 등록을 등록으로 제대로 예측한 경우는 전체 관찰치 중 65.0%로 나타났다.

3.4 신경망분석

신경망 분석이란 인간이 학습(learning)을 통해서 다음의 행동을 행하는 원리를 기계 또는 컴퓨터에 적용시킨 것으로서 기계 또는 컴퓨터에 훈련용 데이터를 이용하여 가장 최적의 결과를 학습시키고 새로운 데이터 또는 상황에 그 학습의 결과를 응용하여 예상결과를 도출하게 하는 분석기법이다.

분석 데이터에 적용한 신경망 모형은 은닉층(Hidden layer)이 1개인 다중퍼셉트론(Multi-layer perceptron) 알고리즘을 적용하였으며, 신경망 분석의 실행결과 설명변수들에 대한 중요도를 살펴보면 아래 <표 8>과 같다.

<표 8> 신경망 분석 결과

설명변수	중요도
대학명	0.33746
학과계열	0.18871
지역	0.17221
탐구영역	0.16072

위의 <표 8>에서 알 수 있는 바와 같이, 대학명, 학과계열, 지역, 탐구영역 등이 중요한 변수로 선택된 것을 알 수 있는데, 이와 같은 결과는 로지스틱 회귀모형의 결과와 비교해 보면 다소간의 차이가 있음을 알 수 있다. 또한 신경망 분석의 결과를 좀 더 쉽게 이해하기 위해서 의사결정나무 모형의 형태로 표현하여 이탈가능성이 높은 집단을 선별해 보면, 예대을 지원하면서 사회탐구를 선택했고 지역이 부산인 경우, 문대를 지원하면서 내신성적이 18.612이하인 경우, 지역이 경남이면서 내신성적이 34.037이하인 경우, 멀대를 지원하면서 내신성적이 20.208이하인 경우, 이들 집단에 대한 이탈율이 모두 높게 나타났다.

위에서 적용한 모형에 대한 오분류율의 결과를 정리한 것이 아래 <표 9>와 같다.

<표 9> 신경망 모형에 의한 오분류표

실제 / 예측		실제범주		계
		등록	이탈	
예측범주	등록	0.46	0.09	0.55
	이탈	0.28	0.17	0.45
계		0.74	0.26	1.00

위의 결과를 보면 구축된 모형은 실제 이탈을 이탈로, 실제 등록을 등록으로 제대로 예측한 경우는 전체 관찰치 중 63.0%로 나타났다.

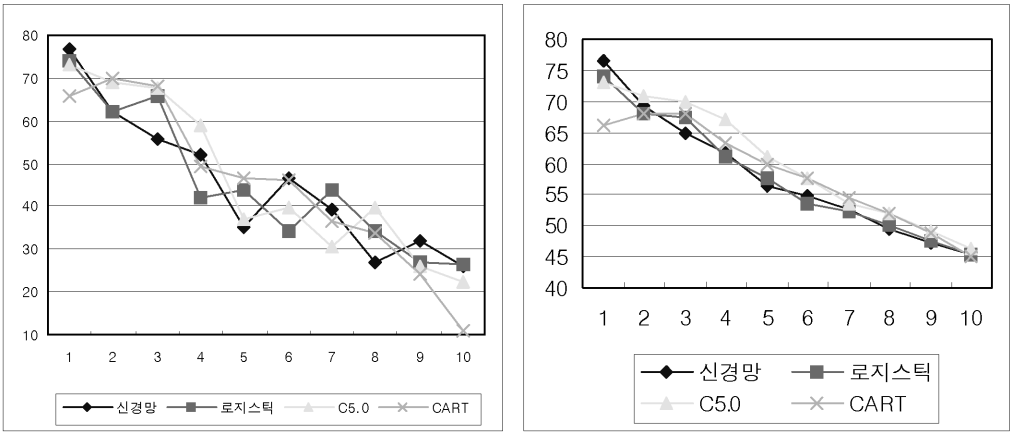
4. 이익도표에 의한 모형평가

이 절에서는 이익도표를 이용하여 4개의 예측모형을 비교하여 보다 나은 모형을 선택하고자 한다. 이익도표에서 사용할 통계량은 다음과 같이 반응퍼센트(% Response)와 리프트(Lift)로 다음과 같이 정의된다.

$$\text{반응퍼센트} = \frac{\text{해당집단에서 이탈학생의빈도}}{\text{해당집단에서 전체빈도}} \times 100, \tag{4}$$

$$\text{리프트} = \frac{\text{해당집단에서 이탈학생의빈도}}{\text{전체집단에서이탈학생의 빈도}} \times \frac{1}{\text{집단개수}}. \tag{5}$$

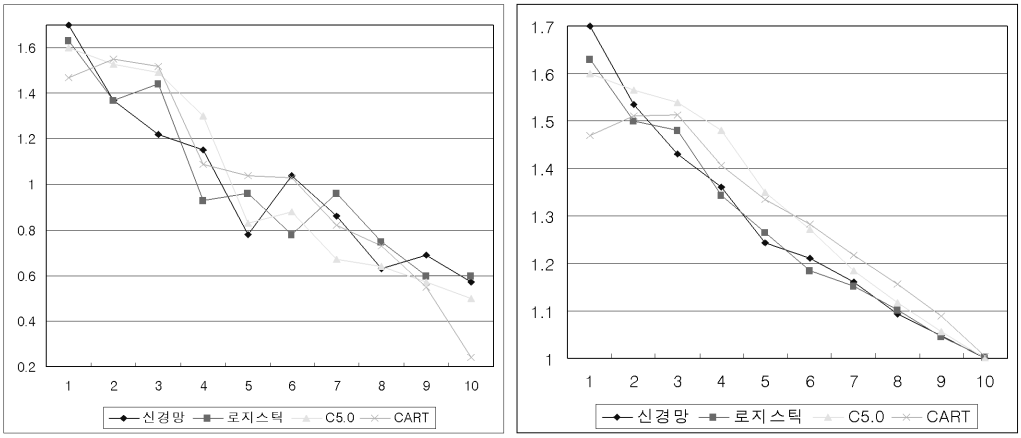
여기서 편의상 4개의 예측모형에서 집단의 구분은 편의상 10등분으로 하였다. 아래 <그림 1>은 각 예측모형에 대한 반응 퍼센트(누적 반응 퍼센트)를 보여주고 있다.



<그림 1> 4개의 예측모형에 대한 반응 퍼센트

위의 <그림 1>에 따르면 이탈가능성이 높은 상위 30%의 최초합격자들을 집중관리를 한다면 C5.0 모형의 경우, 약 70%정도, CART 모형의 경우 약 68% 정도, 로지스틱 회귀모형의 경우 약 67.5%, 신경망 모형의 경우 약 64% 정도의 이탈 학생을 방지할 수 있다는 것을 알 수 있다. 이런 경향은 이탈가능성이 높은 상위 10%를 제외한 20%, 30%, 40%, 50%에 대해서도 C5.0 모형이 다른 예측모형에 비해서 높은 성능을 보임을 알 수 있다.

아래 <그림 2>는 4개의 예측모형에 대한 리프트(누적 리프트)를 보여주고 있다.



<그림 2> 4개의 예측모형에 대한 리프트

리프트 도표의 특징은 각 집단 내에서 이탈율을 전체 이탈율에 비교해서 볼 수 있다는 점이다. 위의 <그림 2>에서 알 수 있는 바와 같이, 이탈가능성이 높은 상위 10%의 집단에 대해서는 신경망 모형이 전체 이탈율보다 약 1.7배 정도로 가장 높은데 반해, 이탈가능성이 높은 상위 20%, 30%, 40%, 50%에 대해서는 C5.0 모형이 다른 모형에 비해서 가장 높은 것을 알 수 있다. 또한 C5.0 모형이 다른 모형에 비해서 상대적으로 이탈이 확실한 집단과 그렇지 못한 집단에 대한 구분은 뚜렷하게 구분해 주고 있음을 알 수 있다.

따라서 본 연구에서는 이익도표에 의한 모형평가를 통해서 C5.0 모형으로 이탈학생들을 예측하는 모형으로 추천하고자 한다.

5. 결론

앞의 4절의 결과에 따라서 이익도표에 의한 모형평가를 통해 4개의 예측모형 중에서 보다 나은 모형으로 C5.0에 의한 예측모형이 선택되어졌음을 알 수 있다. 이상의 분석과정을 통해 얻어진 결과들을 실제 입시문제에 적용하기 위해서는 C5.0 예측모형에 의한 이탈가능성이 높을 것으로 예상되는 집단에 대한 리스트를 파악하는 것이 중요하며 이러한 리스트를 이용해서 이탈방지를 위한 홍보 전략을 수립할 필요가 있다.

참고문헌

1. 김순귀, 정동빈, 박영술(2003). SPSS를 활용한 로지스틱 회귀모형의 이해와 응용, SPSS 아카데미, 서울
2. 최종후 · 한상태 · 강현철 · 김은석(1998). AnswerTree를 이용한 데이터마이닝 의사결정나무분석, SPSS 아카데미, 서울.
3. 허명회, 이용구(2003). 데이터마이닝 모델링과 사례, SPSS 아카데미, 서울
4. 허준, 정규상, 허수희, 최희경, 정성원(2003). Clementine 7 매뉴얼, SPSS 아카데미, 서울
5. 허준, 최병주, 정성원(2001). 클레멘타인을 이용한 데이터 마이닝 입문, SPSS 아카 데미, 서울
6. Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone(1984). *Classification and regression trees*, Belmont : Wadsworth.
7. Kass, G.(1980). An exploratory technique for investigating large quantities of categorical data, *Applied Statistics*, 29(2), 119-127.
8. Quinlan, J. R.(1993). *C4.5 Programs for machine learning*. San Mateo : Morgan Kaufmann

[2006년 11월 접수, 2006년 12월 채택]