디지털산업정보학회 논문지 제19권 제4호-2023년 12월
http://dx.doi.org/10.17662/ksdim.2023.19.4.137

시프트 시그모이드 분류함수를 가진 로지스틱 회귀를 이용한 신입생 중도탈락 예측모델 연구*

김 동 형**

A Study of Freshman Dropout Prediction Model Using Logistic Regression with Shift-Sigmoid Classification Function

Kim Donghyung

〈Abstract〉

The dropout of university freshmen is a very important issue in the financial problems of universities. Moreover, the dropout rate is one of the important indicators among the external evaluation items of universities. Therefore, universities need to predict dropout students in advance and apply various dropout prevention programs targeting them. This paper proposes a method to predict such dropout students in advance. This paper is about a method for predicting dropout students. It proposes a method to select dropouts by applying logistic regression using a shift sigmoid classification function using only quantitative data from the first semester of the first year, which most universities have. It is based on logistic regression and can select the number of prediction subjects and prediction accuracy by using the shift sigmoid function as an classification function.

As a result of the experiment, when the proposed algorithm was applied, the number of predicted dropout subjects varied from 100% to 20% compared to the actual number of dropout subjects, and it was found to have a prediction accuracy of 75% to 98%.

Key Words : Prediction Dropout of Freshman, Logistic Regression, Shift-Sigmoid Classification Function, Prediction Model, Student Dropout Rate in University

Ⅰ. 서론

대학정보공시에 따르면 지난 3년 (기준연도 2020년~2022년) 전국 대학의 신입생 평균 중도탈락(휴학, 자퇴, 제적 등) 비율은 각각 7.39%, 8.00%, 9.05%로 해마다 증가하고 있다[1]. 학생들의 중도탈락은 대학의 재정적 관점에서도 중요한 이슈이지만, 무엇보다 대학을 대상으로 한 다양한 외부 평가의 중요한 지표로 사용되기 때문에 이에 대한 적극적인 관리가 필요하

* 본 논문은 2023년도 2학기 한양여자대학교 교내연구비에 의하여 연구됨
** 한양여자대학교 소프트웨어융합과 부교수(단독저자)

다. 이러한 이유로 중도탈락 학생들을 사전에 선별하고 대상 학생들에 대한 선제적 학생 관리를 도모하여 대학의 중도탈락률을 줄이기 위한 다양한 연구들이 진행되어 왔다.

이러한 연구는 크게 중도탈락 현황 및 현상 자체에 대한 연구[2, 3], 중도탈락 예측 모델을 생성하기 위한 데이터의 수집에 관한 분야, 그리고 이렇게 수집된 데이터를 활용하여 예측모델을 생성하는 분야로 구분될 수 있다. 데이터 수집과 관련하여 임종민의 연구에서는 학생들의 중도탈락을 사전에 예측 가능할 수 있도록 학생 특성, 대학 특성, 학업 여건, 자퇴 성향 등을 기반으로 대학생활적응능력에 대한 설문문항을 개발하여 검증하였다[4]. 수집된 데이터 세트 중 결측 데이터의 처리방법을 제안하여 기존에 단순히 평균값 등으로 결측 데이터를 추정한 방법과 비교하여 동일 예측모델에서 높은 성능을 가짐을 보였다[5]. 이지은의 연구에서는 사이버 대학의 중도탈락과 관련한 다양한 예측 변수 중 연관성이 큰 예측변수를 추출하는 연구를 수행하였으며, 실험결과 다양한 예측 변수 중 수강차시(진도율), 이수학점, 평점, 휴학 횟수가 중도탈락에 유의미한 예측변수인 것으로 나타났다[6]. 그 밖에 대학생활적응 검사결과의 항목을 의사결정나무에 적용하여 중도탈락 영향 요인을 탐색한 방법 등이 연구되었다[7].

중도탈락 예측모델 자체를 생성하는 연구로 중도탈락자를 대상으로 한 중도탈락 사유와 의도에 관한 설문 텍스트로부터 텍스트를 추출 및 분석하였으며, 추출된 데이터를 4가지 분류 알고리즘에 각각 적용하여 예측 모형을 개발하고 평가하였다[8]. 또한 학업성과와 인구 통계학적 지표 사이의 상관 관계를 분석하고 이를 측정하였으며, KNN(K-Nearest Neighbor), NB(Naive Bayes), 그리고 DT(Decision Tree) 분류기법을 이용하여 최적의 모델링을 수행하는 방법을 제안되었다[9]. 그 밖에 선행학습자의 학습결과를 토대로 원격대학의 중도탈락 예측시스템을 구축하기 위해 필요한 방안 등이 연구되었다[10].

제안하는 논문은 추가적인 설문 또는 검사 없이 1학년 1학기에 이수과정에서 자연스럽게 취득 가능한 정량데이터를 기반으로 신입생의 중도탈락을 예측하는 로지스틱 회귀 기반의 중도탈락 예측 방법을 제안한다. 예측과정에 다양한 크기로 시프트된 시그모이드 분류 함수를 사용함으로써 중도탈락 예측 대상자 수와 예측 정확도를 선택할 수 있다.

논문의 구성은 2장에서 관련이론으로서 제안한 예측 방법의 기반이 되는 로지스틱 회귀 방법과 중도탈락 예측에 관한 대표적인 방법들을 소개한다. 3장에서 제안하는 알고리즘을 각 단계별로 기술하고, 4장에서 다양한 환경에서의 중도탈락 대상자 수와 예측 정확도를 비교한 후, 마지막 절에서 결론을 맺는다.

Ⅱ. 관련 이론

본 절에서는 제안하는 예측모델 생성의 기반이 되는 로지스틱 회귀방법과 기존 중도탈락 예측 모델 기법을 소개한다.

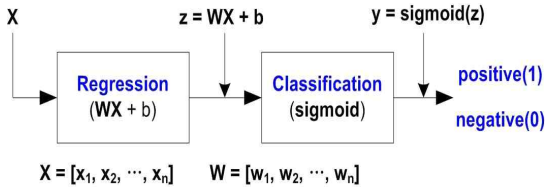
2.1 로지스틱 회귀(Logistic Regression)

2.1.1. 로지스틱 회귀의 동작 매커니즘

분류(classification)란 임의의 입력 특성값들의 조합이 어떤 종류로 예측될 수 있는지를 예측하는 것을 말한다. 예를 들어 메일의 제목, 내용, 특정 문자열의 존재 여부 등을 기반으로 스팸 메일인지 아닌지를 분류하는 것이나, 피검사로 나타난 다양한 수치들의 조합으로 당뇨, 고지혈증, 정상 등을 분류하는 예도 이에 해당한다. 이러한 분류를 수행하는데 대표적으로 적용될 수 있는 방법이 로지스틱 회귀이다.

로지스틱 회귀는 <그림 1>과 같이 먼저 선형회귀

(linear regression)를 이용하여 입력 특성 데이터의 분포를 나타내는 최적의 직선(regression)을 구한 이후 그 직선을 기준으로 양성(1)과 음성(0)을 분류해주는 알고리즘이다. 각 과정을 보다 자세하게 기술하면 먼저 회귀 단계의 결과는 식 (1)과 같이 각 입력 특성 값에 가중치를 곱한 값들의 합 즉, 가중합(weighted sum)으로 표현될 수 있다.



<그림 1> 로지스틱 회귀의 블록다이어그램

$$z = WX + b \quad (1)$$

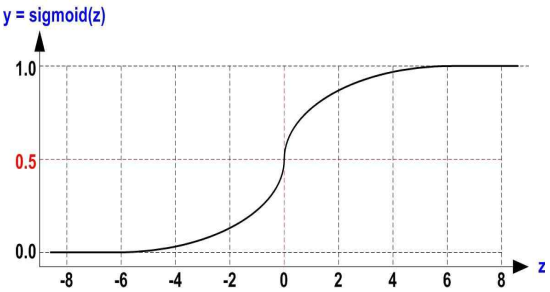
$where,$

$X = [x_1, x_2, \ldots, x_n],\ W = [w_1, w_2, \ldots, w_n]$

여기서 X는 입력특성 벡터, W는 가중벡터, 그리고 b는 바이어스(bias) 값을 의미한다. 식 (1)의 경우 z의 값은 -∞~∞ 범위의 값을 가진다. 한편 분류를 위해서는 z값을 0~1로 매핑 할 분류함수(classification function)가 필요하다. 로지스틱 회귀방법에서는 이러한 분류함수로 시그모이드(sigmoid) 함수를 사용하며 식 (2)와 같다.

$$y = sigmoid(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(WX+b)}} \quad (2)$$

<그림 2>는 선형회귀의 출력값(z)에 따른 로지스틱 회귀의 최종 출력을 나타내는 시그모이드 분류함수의 출력(y)을 그래프로 나타낸 것이다. 여기에서 볼 수 있는 바와 같이 로지스틱 회귀의 최종결과로 나온 y값은 $0 \sim 1$의 범위 값을 가지며, $y \geq 0.5$인 경우 양성(1), $y < 0.5$인 경우 음성(0)이라고 분류하는 것이다.



<그림 2> 시그모이드 분류 함수 그래프

2.1.2. 로지스틱 회귀의 손실함수

논리적인 분류값(양성(1), 음성(0))을 가지는 로지스틱 회귀는 출력값으로 연속적인 값을 가지는 선형회귀와는 달리 상관엔트로피(cross-entropy)를 손실함수로 사용하며 식 (3)은 이진분류를 수행하는 로지스틱 회귀에서의 손실함수를 나타낸다.

$$E(W, b) = -\sum_{i=1}^{n} [t_i \log(y_i) + (1-t_i)\log(1-y_i)] \quad (3)$$

$where,$

$$y = \frac{1}{1+e^{-(WX+b)}},\ t_i = 0 \text{ or } 1$$

최적의 분류 모델은 식 (4)와 같이 손실함수에 포함된 가중치 벡터 W와 바이어스 b를 줄이는 방향으로 학습함으로써 구할 수 있다.

$$W = W - \alpha \frac{\partial E(W,b)}{\partial W} \quad (4)$$

$$b = b - \alpha \frac{\partial E(W,b)}{\partial b}$$

여기서 α는 학습 비율(learning rate)로 학습에 따른 가중치와 바이어스 값의 증감의 정도를 나타낸다.

2.2 중도탈락예측에 관한 이전 연구

중도탈락에 관련한 연구는 크게 중도탈락 예측을 위한 변수 추출 또는 설문문항 개발을 개발하는 분야와 다양한 인공지능 분류기법을 활용하여 중도탈락 예측 모델을 생성하는 분야로 나뉠 수 있다. 여기에서는 실제 중도탈락 예측모델을 생성하는 대표적인 방법으로 분류기법을 활용한 방법과 앙상블 분류법을 활용한 중도탈락 예측모델 생성 기법을 소개한다.

2.2.1. 분류기법을 활용한 중도탈락 예측

정선호가 제안한 중도탈락 예측모형기법[8]은 2017년부터 2021년까지의 특정 대학 중도 탈락자를 대상으로 한 설문과 2019학년도에서 2020학년도에 재학중인 학생들을 대상으로 한 설문 중에서 중도 탈락의 사유와 의도에 대한 텍스트를 분석 데이터로 사용하였다.

STEP1. 데이터 전처리

- 주어진 텍스트에서 불필요한 부분을 제거하고 단어를 분리하는 처리를 수행
- 연속된 문자열로 표현된 텍스트 데이터를 의미 표현의 기본 단위인 토큰으로 나누어주는 토큰화 수행

STEP2. 특성 추출

- TF-IDF(Term Frequency-Inverse Document Frequency, 단어빈도, 문서역빈도) 모델을 사용하여 대상 텍스트를 텍스트의 특성을 나타내는 수치 형태로 변환.

STEP3. 데이터 학습

- STEP2에서 TF-IDF로 벡터화하여 수치화한 데이터 세트를 각기 다른 알고리즘인 NB, LR, DT, RF를 사용하여 모형을 학습함

이후 4개의 분류 알고리즘 (NB(Naive Bayes), LR(Logistic Regression), DR(Decision Tree), RF(Random Forest))을 적용하여 예측모델을 생성하여 비교하였다. 예측모델 생성 단계는 다음과 같이 크게 3단계로 구성되어 있다.

이상의 학습을 통한 결과로 학습용 데이터에 대한 정확도는 DT와 RF가 높은 것으로 나타났으며, 예측 정확도는 NB와 LR 알고리즘이 높은 것으로 나타났다.

2.2.2 앙상블 분류법을 이용한 중도탈락 예측

앙상블 분류법을 이용하여 중도탈락 예측 모델링 기법은 중도탈락을 예측하기 위해 KNN, NB, 그리고 DT의 세 가지 분류기법을 사용하여 학업 성과와 인구 통계학적 지표 사이의 상관관계를 분석하고 측정한다[9]. 이 방법 역시 예측모델 생성 단계가 다음과 같이 크게 3단계로 구성되어 있다.

STEP1. 데이터 구성 및 상관속성 추출

- 대학의 학사정보시스템에서 학생 중도탈락과 관련된 데이터를 추출하여 학습용 데이터 구성
- 앙상블 배깅 트리 방법을 사용하여 특징 선택을 수행하여 중도탈락을 예측할 수 있는 상관 속성 추출

STEP2. 모델 샘플 도출

- 구성된 데이터를 사용하여 인공신경망(ANN), 의사결정트리(DT), 베이지안 네트워크(BN) 등의 머신러닝 알고리즘을 기반으로 구성된 예측모델을 훈련시켜 모델 샘플 도출

STEP3. 예측모델 최적화

- 3단계 테스트 데이터 세트를 생성된 예측 모델에 입력, 앙상블 결정트리를 적용하여 예측 모델 정확도 최적화

이상의 과정을 이용하여 실제 데이터에 적용한 결과 앙상블 분류기를 이용한 예측 모델이 가장 높은 정확도를 가지는 것으로 나타났고, NB가 두 번째로 높은 정확도를 보이는 것으로 나타났다.

Ⅲ. 제안하는 알고리즘

3.1 예측 대상

먼저 제안하는 알고리즘이 예측하고자 하는 대상을 명확히 할 필요가 있다. 중도탈락 대상을 예측하는데 있어서 <표 1>과 같이 4가지의 경우가 존재한다.

<표 1> 중도탈락 예측 매트릭

예측값 / 실제값	중도탈락	재학
중도탈락	**CASE 1** 중도탈락→중도탈락	**CASE 2** 중도탈락→재학
재학	**CASE 3** 재학→중도탈락	**CASE 4** 재학→재학
합계	SumOfPred(1,3)	SumOfPred(2,4)

CASE 1과 CASE 2의 경우 실제 중도탈락 대상 학생을 각각 중도탈락이라고 옳게 예측한 경우와 그렇지 못한 경우이다. 유사하게 CASE 3과 CASE 4는 실제 재학 중인 학생을 각각 중도탈락이라고 잘못 예측한 경우와 바르게 예측한 경우를 의미한다. 제안하는 알고리즘은 이 4가지 경우 중에서 중도탈락이라고 예측한 전체 대상(SumOfPred(1,3)) 대비 CASE 1의 비율을 높이는 것을 목표로 한다. 이 경우 명확한 중도탈락 대상자를 선별함으로써 보다 집중적으로 중도탈락 방지를 위한 후속 조치를 취할 수 있게 된다.

<표 2> 중도탈락 예측 매트릭

No.	항목	값의 범위
1	중간고사점수	0 ~ 30
2	기말고사점수	0 ~ 40
3	출석점수	0 ~ 30
4	평점평균	0 ~ 4.5
5	백분위점수	0 ~ 100
6	상담건수	0 ~ 10
7	비교과건수	0 ~ 10
8	장학금수혜액	0 ~ 5,000,000

3.2 데이터의 선정

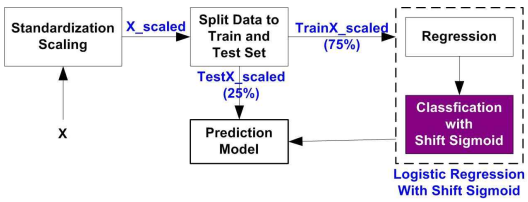
일반적으로 중도탈락 예측에 사용되는 데이터는 크게 분류 데이터와 정량데이터로 나뉠 수 있다. 분류 데이터는 비선형성 데이터로 입학전형(수시1, 수시2, 정시 등), 주소지(서울, 경기, 인천 등), 고교계열(인문계, 특성화고, 검정고시 등) 등이 있다. 정량 데이터의 경우 선형적으로 수치화된 데이터를 의미하며 평점, 상담회수, 시험성적, 장학금 수혜금액 등이 여기에 포함된다. 학생들의 중도탈락을 예측하기 위해 개발된 설문 결과 또한 정량데이터의 한 종류라고 할 수 있다.

이들 데이터 중에서 분류데이터의 경우 각 분류별 데이터의 수가 고르지 않아 데이터의 편향성이 생길 수 있고 충분한 데이터를 확보하는데 어려움이 있을

수 있다. 또한 중도탈락 예측을 위해 특화된 설문 결과 등의 정량 데이터는 '휴학 또는 자퇴할 의향이 있느냐'와 같은 직관적인 질문에 대한 대답 등이 포함되어 있어 예측에는 유리하지만, 설문당 추가적인 비용이 발생하여 해당 데이터가 없는 대학이 대부분이다. 이러한 이유로 제안하는 방법은 <표 2>와 같이 1학년 1학기에 자연스럽게 취득될 수 있는 정량데이터만을 중도탈락 예측모델에 사용함으로써 결측 데이터를 최소화하며, 1학년 1학기를 이수한 학생들은 쉽게 예측모델을 적용할 수 있도록 하였다.

3.3 중도탈락 예측 수행 과정

제안하는 중도탈락 예측모델의 생성과정은 <그림 3>과 같이 구성된다.



<그림 3> 제안하는 알고리즘의 블록 다이어그램

먼저 입력 데이터는 스케일링(scaling)을 통해 동일 범위내로 매핑을 수행한다. 입력 데이터 벡터 X는 앞서 <표 2>에서 볼 수 있는 바와 같이 서로 다른 범위 값을 가지며, 특히 장학금의 경우 가질 수 있는 값의 범위가 다른 입력 특성과 큰 차이를 보인다. 이 경우 예측 모델이 특정 입력 특성에 지나치게 편중되는 문제가 발생할 수 있다. 따라서 전체 입력 특성의 범위를 동일하게 조정하는 과정이 필요한데 이것이 데이터 스케일링이며 제안하는 방법에서는 식 (5)의 표준 스케일링(standardization scaling) 방법을 사용하였다.

$$x_{scaled} = \frac{x - mean(x)}{std(x)} \tag{5}$$

여기서 $mean(x)$와 $std(x)$는 각각 입력 특성 x값들에 대한 평균 및 표준편차를 의미한다. <표 3>은 표준 스케일링 전의 데이터와 표준 스케일링 후의 데이터를 비교한 예이다.

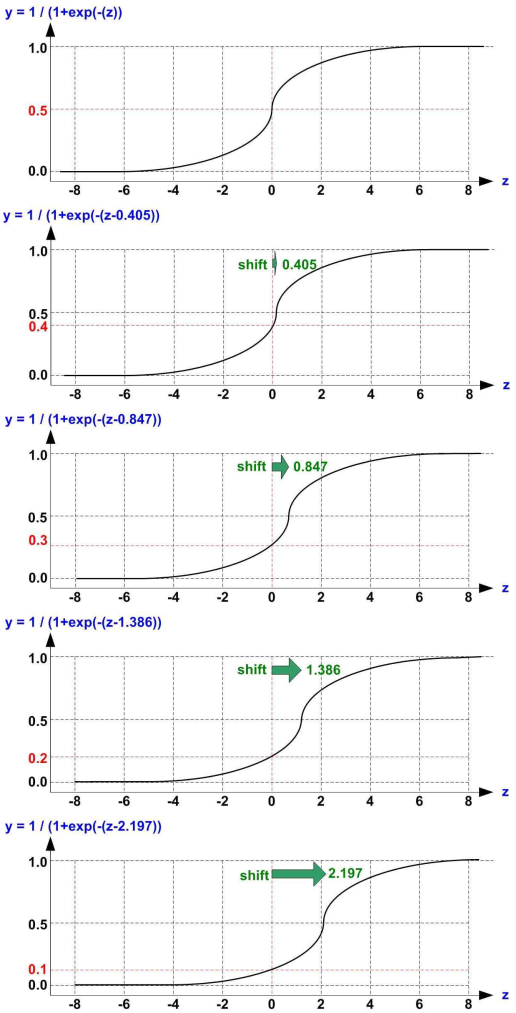
<표 3> 입력 데이터의 표준 스케일링 값 비교 예

No.	X	X_scaled
1	2.2000×10^{01}	-0.02777803
2	3.0000×10^{01}	0.25018308
3	2.7000×10^{01}	0.18848606
4	2.8000×10^{00}	-0.00930595
5	7.9000×10^{01}	0.13920741
6	1.0000×10^{00}	-0.42832932
7	3.0000×10^{00}	0.81360313
8	2.5146×10^{05}	-0.83104737

표준 스케일링 이후 과적합(overfitting)을 방지하기 위해서 전체데이터는 학습용 데이터와 검증용 데이터로 나눈 후 학습용 데이터만을 이용하여 로지스틱 회귀를 적용한다. 제안하는 방법에서는 학습용 데이터와 검증용 데이터를 각각 75%와 25%로 할당하였다.

제안하는 알고리즘에서는 로지스틱 회귀를 적용하는 과정에서 분류함수로 시프트(shift) 된 시그모이드 함수를 사용한다.

<그림 4>는 시그모이드 함수를 각각 0, 0.405, 0.847, 1.386, 2.197 시프트 시킨 분류함수의 모양을 나타낸다. 그림에서도 볼 수 있는 바와 같이 시프트의 정도에 따라 예측결과는 서로 다르게 나올 것이다. 예를 들어 시프트가 되지 않은 기본 시그모이드 분류함수의 경우 중도탈락 확률이 0.5이상의 경우 중도탈락 대상자로 예측한다. 반면 마지막 시프트 시그모이드 분류함수(<그림 4> (e))를 사용한 경우 시프트가 되지 않은 시그모이드 기준 중도탈락 확률이 0.9 이



<그림 4> 시프트가 각각 (a) shift=0 (b) shift=0.405 (c) shift=0.847 (d) shift=1.386 (e) shift=2.197 된 시그모이드 분류함수의 그래프

상의 경우(<그림 4> (a)의 z(y=0.9)의 위치는 <그림 4> (e)의 z(y=0.5)에 해당) 중도탈락 대상자로 예측한다. 즉 시프트가 많이 될수록 중도탈락 대상자의 수는 줄어들겠지만 예측 정확도는 당연히 올라갈 것이다.

IV. 실험 및 분석 결과

중도탈락 대상자를 예측하기 위해서 1학년 여름방학 ~ 1학년 겨울방학사이에 중도 탈락한 2,207 학생에 대한 실제 데이터를 사용하였으며, 데이터 편향성이 생기지 않도록 중도탈락하지 않은 일반 재학생 2,000명의 데이터를 임의로 선정하여 전체 4,207명의 데이터를 사용하였다. <표 4>는 이상의 데이터 분포를 나타낸다.

<표 4> 실험에 사용된 데이터의 분포

구분	데이터수		합계
재학	2,000		2,000
중도탈락	휴학	982	2,207
	자퇴	933	
	제적	292	
합계			4,207

앞장에서 기술한 바와 같이 모든 데이터는 입력 특성 편향성을 방지하기 위해 표준 스케일링를 수행하였으며, 과적합을 방지하기 위해 전체 데이터를 학습용 데이터(3,155 (75%))와 검증용 데이터(1,052 (25%))로 각각 나누었다. 이후 학습용 데이터를 로지스틱 회귀에 적용하였다.

<표 5> ~ <표 9>는 학습용 데이터에 대해서 시프트 시그모이드 분류함수에 따른 실제값과 예측값에 대한 매트릭(matric)을 나타내며, 예측 정확도는 3.1절에서 기술한 바와 같이 예측 대상인 실제 중도탈락 학생을 중도탈락을 옳게 예측한 확률을 나타낸다.

먼저 분류함수 시그모이드를 시프트하지 않은 <표 5>의 경우 학습용 데이터인 3,155명 대상 학생 중 1,557명을 중도탈락 학생으로 예측했고, 이중 실제 중도탈락 학생은 1,166명이며, 74.89%의 예측 정확도(실제:중도탈락→예측:중도탈락)를 가졌다. <표 7>과 같이 시그모이드를 0.847만큼 시프트한 분류함수를 사용하는 경우 전체 중 641명의 학생을 중도탈락 할 것

<표 5> **학습용** 데이터에 shift=0.0인 시그모이드 분류함수를 사용한 경우의 중도탈락 예측자 수 및 예측정확도

shift = 0.0		예측값		
		중도탈락	재학	합계
실제값	중도탈락	1,166	502	1,668
	재학	391	1,096	1,487
	합계	1,557	1,598	3,155
예측정확도		**74.89 %**		

<표 6> **학습용** 데이터에 shift=0.405인 시그모이드 분류함수를 사용한 경우의 중도탈락 예측자 수 및 예측정확도

shift = 0.0		예측값		
		중도탈락	재학	합계
실제값	중도탈락	881	787	1,668
	재학	136	1,351	1,487
	합계	1,017	2,138	3,155
예측정확도		**86.63 %**		

<표 7> **학습용** 데이터에 shift=0.847인 시그모이드 분류함수를 사용한 경우의 중도탈락 예측자 수 및 예측정확도

shift = 0.0		예측값		
		중도탈락	재학	합계
실제값	중도탈락	588	1,080	1,668
	재학	53	1,434	1,487
	합계	641	2,514	3,155
예측정확도		**91.73 %**		

<표 8> **학습용** 데이터에 shift=1.386인 시그모이드 분류함수를 사용한 경우의 중도탈락 예측자 수 및 예측정확도

shift = 0.0		예측값		
		중도탈락	재학	합계
실제값	중도탈락	436	1,232	1,668
	재학	21	1,466	1,487
	합계	457	2,698	3,155
예측정확도		**95.40 %**		

<표 9> **학습용** 데이터에 shift=2.197인 시그모이드 분류함수를 사용한 경우의 중도탈락 예측자 수 및 예측정확도

shift = 0.0		예측값		
		중도탈락	재학	합계
실제값	중도탈락	312	1,356	1,668
	재학	10	1,477	1,487
	합계	322	2,833	3,155
예측정확도		**96.89 %**		

이라고 예측하였고, 이 중 588명이 실제 중도탈락하여 91.73%의 예측 정확도를 가졌다. 마지막으로 <표 9>와 같이 2.197만큼 시프트를 하는 경우의 예측 정확도는 95.40%까지 증가하지만 중도탈락 예측 학생수는 보다 줄어들게 된다. 관리 대상을 소수화하여 집중 관리하고자 하는 경우 분류함수의 시프트의 크기를 증가시키는 것이 유리하겠지만 91.73%의 예측 정확도를 가지는 <표 6>의 결과도 충분하리라 판단된다.

이상의 데이터는 중도 탈락 모델 생성과정에서 사용된 학습용 데이터에 대한 예측 정확도를 나타낸다. 과적합 여부를 확인하기 위해 학습과정에서 사용되지 않은 25%의 검증데이터에 동일한 시프트 분류함수를 적용하여 확인한 예측 정확도는 <표 10> ~ <표 14>와 같다.

결과에서 볼 수 있는 바와 같이 시프트가 0.0, 0.405, 0.847, 1.386, 2.197의 경우 각각 75.46%, 88.78%, 93.07%, 97.12%, 98.02%의 예측 정확도로 오히려 학습용 데이터에서 보다 높은 예측 정확도를 가짐으로써 예측모델에 과적합은 발생하지 않았음을 알 수 있다. 이상의 결과를 통해 중도탈락 예측 대상자에 대한 향후 관리방안의 특징, 성격 등에 따라 시그모이드 분류 함수의 시프트 크기를 정할 수 있을 것이다. 예를 들어 간단한 지도교수 상담이나 추가 지원 등의 경우 예측 정확도 보다는 대상자를 늘리는 것이 유리할 것이다.

하지만 소수의 인원에 대해서 장기간에 걸친 집중 프로그램을 운영하고자 하는 경우 중도탈락이 명확한 소수의 학생을 대상을 진행하는 것이 효율적일 것이다. 즉, 대상자의 수와 예측 정확도는 상호 trade-off 관계를 가지고 있으며, 이 둘 사이의 가중치를 시그모이드 분류함수의 시프트 양으로 조절할 수 있는 것이다.

<표 10> **검증** 데이터에 shift=0.0인 시그모이드 분류함수를 사용한 경우의 중도탈락 예측자 수 및 예측정확도

shift = 0.0		예측값		
		중도탈락	재학	합계
실제값	중도탈락	372	167	539
	재학	121	392	513
	합계	493	559	1,052
예측정확도		**75.46 %**		

<표 11> **검증** 데이터에 shift=0.405인 시그모이드 분류함수를 사용한 경우의 중도탈락 예측자 수 및 예측정확도

shift = 0.0		예측값		
		중도탈락	재학	합계
실제값	중도탈락	277	262	539
	재학	35	478	513
	합계	312	740	1,052
예측정확도		**88.78 %**		

<표 12> **검증** 데이터에 shift=0.847인 시그모이드 분류함수를 사용한 경우의 중도탈락 예측자 수 및 예측정확도

shift = 0.0		예측값		
		중도탈락	재학	합계
실제값	중도탈락	188	351	539
	재학	14	499	513
	합계	202	850	1,052
예측정확도		**93.07 %**		

<표 13> **검증** 데이터에 shift=1.386인 시그모이드 분류함수를 사용한 경우의 중도탈락 예측자 수 및 예측정확도

shift = 0.0		예측값		
		중도탈락	재학	합계
실제값	중도탈락	135	404	539
	재학	4	509	513
	합계	139	913	1,052
예측정확도		**97.12 %**		

<표 14> **검증** 데이터에 shift=2.197인 시그모이드 분류함수를 사용한 경우의 중도탈락 예측자 수 및 예측정확도

shift = 0.0		예측값		
		중도탈락	재학	합계
실제값	중도탈락	99	440	539
	재학	2	511	513
	합계	101	951	1,052
예측정확도		**98.02 %**		

Ⅴ. 결론 및 향후연구

학령인구 감소로 신입학은 물론 편입 등의 기회가 확대됨에 따라 대학의 중도 탈락률은 지속적으로 증가하고 있는 추세이다. 대학입장에서는 이러한 중도탈락학생들을 사전에 선별하고, 이들에 대해서 중도탈락 방지를 위한 다양한 관리방안을 갖추는 것이 무엇보다도 중요해지고 있다. 본 논문은 중도탈락 학생을 예측하는 방안에 관한 것으로 대부분이 대학이 가지고 있는 1학년 1학기 정량 데이터만으로 시프트 시그모이드 분류함수를 사용한 로지스틱 회귀를 적용하여 중도탈락 대상자를 선별하는 방법을 제시하였다. 제안하는 알고리즘에서는 시그모드이 분류함수의 시프트 크기에 따라 중도탈락 예측 대상자 수와 예측정확도를 적절히 결정할 수 있으며, 실험 결과 대상자의 수는 실제 중도탈락 대상자수 대비 100% ~ 20%까지 변화함에 따라 75% ~ 98%의 예측정확도를 가지는 것으로 나타났다.

향후연구로는 1학년 1학기의 정량데이터가 아닌 1학년 1학기 중간고사 이전에 취득할 수 있는 다양한 정량 데이터를 연구하고 이를 중도탈락 예측에 활용함으로써 학기 중 중도탈락 사전 관리가 가능한 예측모델 개발로 연구를 확대하고자 한다.

참고문헌

[1] 대학정보공시, https://www.academyinfo.go.kr/index.do

[2] 정선호, “대학생의 중도탈락 현황연구: P 대학 사례를 중심으로,” 한국안전문화연구원, 융합과 통섭, 제4권, 제3호, 2021, pp.136-145.

[3] 윤소정 · 강승희, “토픽모델링을 활용한 중도탈락 대학생 대상 연구 동향 분석,” 아시아문화학술원, 인문사회 21, 제13권, 제2호, 2022, pp.2803-2814.

[4] 임종민 · 이재혁 · 최윤정 · 김보연 · 박주희, “대학생활 적응력 및 중도탈락 이상기후 예측요인 분석을 위한 설문지 개발,“ 한국고등직업교육학회, 한국고등직업교육학회 논문집, 제20권, 제2호, 2022, pp.1-10.

[5] 박상성, “앙상블 기법을 활용한 대학생 중도탈락 예측 모형 개발,“ 디지털산업정보학회, 디지털산업정보학회논문지, 제17권, 제1호, 2021, pp.109-115.

[6] 이지은, “학생 중도탈락 예측지수에 관한 사후검증 연구,“ 한국빅데이터학회, 한국빅데이터학회지, 제4권, 제2호, 2019, pp.175-183.

[7] 송연주 · 강창완 · 이정희, “의사결정나무를 활용한 대학생 중도탈락 영향 요인 탐색,“ 아시아문화학술원, 인문사회 21, 제13권, 제5호, 2022, pp.2401-2416.

[8] 정선호, “분류 기법을 활용한 대학생 중도탈락 예측모형 개발,“ 한국안전문화연구원, 융합과 통섭, 제5권, 제2호, 2022, pp.174-185.

[9] 정정호 · 정선호, “앙상블 분류법을 이용한 대학생 중도탈락 예측 모델링 : P 대학을 중심으로,“ 한국안전문화연구원, 융합과 통섭, 스페셜호, 2023, pp.66-75.

[10] 황현정 · 박솔잎 · 박형용, “학습결과 분석을 통한 원격대학 중도탈락 예측 시스템 AI 알고리즘 적용방안,“ 한국컴퓨터교육학회, 컴퓨터교육학회논문지, 제24권, 제5호, 2021, pp.63-73.

■ 저자소개 ■

김 동 형
(Kim Donghyung)

2011년 3월~현재 한양여자대학교 소프트웨어융합과 부교수
2008년 9월 한라대학교 정보통신방송공학부 조교수
2007년 9월 한국전자통신연구원(ETRI) 선임연구원
2007년 8월 한양대학교 전자통신전파공학과 (공학박사)
2001년 2월 충북대학교 전자공학과(공학석사)
1999년 2월 충북대학교 전자공학과(공학사)

관심분야 : 영상처리, 멀티미디어통신, 영상압축
E-mail : kimdh@hywoman.ac.kr

논문접수일 : 2023년 12월 02일
수정접수일 : 2023년 12월 10일
게재확정일 : 2023년 12월 15일