한국시뮬레이션학회 논문지
Vol. 32, No. 4, pp. 19-26 (2023. 12)
http://doi.org/10.9709/JKSS.2023.32.4.019
ISSN 1225-5904

머신러닝 기반 대학생 중도 탈락 예측 모델의 성능 비교

정석봉 · 김두연 †

Performance Comparison of Machine Learning based Prediction Models for University Students Dropout

Seok-Bong Jeong · Du-Yon Kim†

ABSTRACT

The increase in the dropout rate of college students nationwide has a serious negative impact on universities and society as well as individual students. In order to proactive identify students at risk of dropout, this study built a decision tree, random forest, logistic regression, and deep learning-based dropout prediction model using academic data that can be easily obtained from each university's academic management system. Their performances were subsequently analyzed and compared. The analysis revealed that while the logistic regression-based prediction model exhibited the highest recall rate, its f-1 value and ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) value were comparatively lower. On the other hand, the random forest-based prediction model demonstrated superior performance across all other metrics except recall value. In addition, in order to assess model performance over distinct prediction periods, we divided these periods into short-term (within one semester), medium-term (within two semesters), and long-term (within three semesters). The results underscored that the long-term prediction yielded the highest predictive efficacy. Through this study, each university is expected to be able to identify students who are expected to be dropped out early, reduce the dropout rate through intensive management, and further contribute to the stabilization of university finances.

Key words : dropout, machine learning, deep learning, dropout prediction

요 약

전국 대학생의 중도 탈락 비율의 증가는 학생 개인 뿐만 아니라 대학과 사회에 심각한 부정적 영향을 끼친다. 본 연구에서는 중도 탈락이 예상되는 학생을 사전에 식별하기 위하여, 각 대학의 학사관리 시스템에서 손쉽게 얻을 수 있는 학적 데이터를 기반으로 머신러닝 분야의 결정트리, 랜덤 포레스트, 로지스틱 회귀 및 딥러닝 기반의 중도 탈락 예측 모델을 구축하고, 그 성능을 비교·분석하였다. 분석 결과 로지스틱 회귀 기반 예측 모델의 재현율이 가장 높았으나 f-1 및 auc 값이 낮은 한계를 보였고, 랜덤 포레스트 기반의 예측 모델의 경우 재현율을 제외한 다른 모든 지표에서 가장 우수한 성능을 보였다. 또한 예측 기간에 따른 예측 모델의 성능을 확인하기 위하여 예측 기간을 단기(1개 학기 이내), 중기(2개 학기 이내) 및 장기(3개 학기 이내)로 나누어 분석해 본 결과, 장기 예측 시 가장 높은 예측력을 보였다. 본 연구를 통해 각 대학은 중도 탈락이 예상되는 학생들을 조기에 식별하고, 이들에 대한 집중 관리를 통해 중도 탈락 비율을 줄이며 나아가 대학 재정 안정화에 기여할 수 있을 것으로 기대된다.

주요어 : 중도 탈락, 머신러닝, 딥러닝, 중도 탈락 예측

Received: 30 August 2023, **Revised**: 16 October 2023, **Accepted**: 27 October 2023

† **Corresponding Author**: Du-yon Kim
E-mail: duyonkim@kiu.ac.kr
Professor, Dept. of Architectural and Civil Engineering, Kyungil University, Kyeongbuk-do, Korea

1. 서론

전국 4년제 대학의 중도 탈락 학생 비율은 2021학년도 기준 9만 7천 326명으로(재적 학생 대비 4.9%)에 달하고 있으며, 그 수와 비율이 꾸준히 증가하고 있다(연합

뉴스, 2022.09.21.). 중도 탈락은 자퇴와 미등록, 미복학 등을 포함하는 것으로, 2021학년도 기준으로 자퇴가 62.4%(6만802명), 미복학 22.6%(2만2천5명), 미등록 10.7%(1만403명) 순으로 나타나고 있다. 중도 탈락은 교육 분야에서 전 세계적으로 우려하는 과제 중 하나로, 강명희 등(Kang et al., 2019)에 따르면 중도 탈락은 학교 부적응, 경제적 부담 등이 주요 원인이기 때문에, 학생 개인과 대학 그리고 사회적으로 부정적인 영향을 끼친다. 학생 개인에게는 비용과 시간적 손실을, 대학에는 재정 악화를, 사회적으로는 인력양성 및 고등교육 접근성 제고라는 정책목표 달성 실패를 초래할 수 있다(Park, 2020).

중도 탈락 관련 선행 연구들은 주로 학생의 중도 탈락 의도에 영향을 미치는 개인적, 심리적 요인들을 규명하거나, 대학 생활 경험과 같은 경험적 요인, 대학의 교육, 연구 활동 및 교육 여건과 관련된 기관적 요인들의 영향력을 파악하고자 했다(Chung et. al, 2015; Kang et. al., 2019; Park, 2020; Lee and Park, 2019; Lee et. al., 2020; Lee and Kang, 2019; Han, 2018). 한편 이러한 연구들은 대부분 특정 기간이나 대상에 대한 조사 기반의 연구로 일반화하여 적용하기 어렵다는 한계를 가지고 있다(Andrade et. al, 2020).

한편 최근에는 머신러닝 기법을 활용하여 중도 탈락을 조기에 예측하고자 하는 노력들이 꾸준히 이루어지고 있는데, 이들 연구들은 데이터 확보를 위해 설문조사를 이용하거나(Jeong, 2022), 연구에 사용한 데이터가 재학생 전체가 아닌 일부에 그쳐 샘플 수가 적거나, 기간을 고려하지 않은 분석 등으로 실제 적용에 한계가 존재한다(Jeong and Park, 2021; Jeong, 2021). 또한, 최근에는 AI기반 중도 탈락 예측의 상용화된 솔루션도 제공되고 있으나, 이 경우도 각 대학의 특성과 보유한 DB 정보를 가공하는 과정 등을 통해 예측 모델을 구축하게 된다.

본 연구에서는 대학의 학사관리 시스템에서 손쉽게 얻을 수 있는 학적 데이터를 기반으로 머신러닝 분야에서 널리 활용되는 결정 트리(Decision Tree, 이후 DT), 랜덤 포레스트(Random Foresest, 이후 RF), 로지스틱 회귀 기법(Logistic Regression, 이후 LR) 뿐만 아니라 최근 널리 활용되는 딥러닝(Deep Learning, 이후 DL) 기반의 예측 모델을 구축하여 그 성능을 비교 분석한다. 더불어 학적 데이터를 기반으로 구현된 예측 모델을 통해 단기(1개 학기 이내), 중기(2개 학기 이내) 및 장기(3개 학기 이내) 예측을 시도해 봄으로써 얼마나 조기에 대학생들의 중도 탈락을 예측할 수 있는지도 함께 분석하고자 한다.

본 연구의 결과를 바탕으로 대학은 보유한 DB 정보를 활용하여 주요 변수를 도출하고 중도 탈락 예측 모델을 구축할 수 있으며, 중도 탈락이 예상되는 학생들을 조기에 발견하여 이들에 대한 사전 관리를 통해 그 비율을 감소시킬 수 있을 것으로 기대된다.

2. 연구 방법

2.1 데이터 수집

본 연구에서는 경상북도에 소재한 A대학교의 2018~2019학년도까지 4학기 동안(계절학기 제외)의 재적생들 전체의 학적 데이터를 수집하였다. A대학교는 모집정원 1600여 명, 교원 250여 명의 중간규모의 사립대학이다.

2018학년도를 데이터 수집의 시작 시점으로 정한 이유는 A대학교 학사관리 시스템의 특성상 해당 시점을 전후하여 가용한 데이터의 종류 및 연속성에 있어 많은 차이가 있기 때문이다. 또한 코로나 팬데믹 기간의 특수성을 고려하여 2020학년도 이후 데이터는 이번 연구에서 활용하지 않았다.

본 연구에서는 특정 시점(학기)의 학생들에 대한 학적 데이터를 기반으로 향후 1개~3개 학기 내에(계절학기 제외) 중도 탈락할 학생들을 예측하는 것을 목적으로, 2개 학년도 4개 학기 동안 총 34,834명의 학생(학기별 누적 샘플 수)들에 대한 학적 데이터를 수집하였다. 이중 다음의 학생들은 분석 대상에서 제외하였다.

- 데이터 추출 시점에 1학년에 재학중인 학생(25.95%)
- 중도 탈락 후, 재입학한 학생(0.27%)
- 데이터 추출 시점에 나이가 40세 이상인 학생(0.84%)

전체 데이터의 25.95%에 해당하는 1학년 학생들의 경우 본 연구의 예측 모델의 주요 변수인 평균 성적, 평균 출석률, 평균 상담 횟수 등의 주요 학적 데이터가 부재하거나 의미를 갖기에 부족하므로 분석 대상에서 제외하였다. 참고로 1학년 학생들의 중도 탈락 예측에는 학적 데이터보다는 입학 관련 자료나 신입생 실태조사 등의 설문조사 결과를 활용한 별도 모델이 적합할 것으로 판단된다. 또한 한번 중도 탈락했다가 재입학한 학생들(전체의 0.27%)의 경우 학적 데이터의 연속성이 떨어지고, 이들의 중도 탈락 여부에는 기존 학생과는 다른 별도의 요인이 작용할 가능성이 크다고 판단해서 분석 대상에서 제외하였다. 전체 데이터의 0.84%에 해당하는 40세 이상의 학생들은 보통 성인 학습자로 간주되며, 학업 행태 및 중도 탈락 사유에 있어 학령기 학생들과 많은 차이점이

예상되어 제외하였다. 이들 학생들을 제외한 분석 데이터는 총 25,567명(학기별 누적 샘플 수)이다.

분석 대상 학생들의 구성을 간략히 살펴보면, 남성이 76.5%, 여성이 23.5%이며, 25세 미만이 89.2%, 25세 이상 30세 미만이 10.5%, 30세 이상이 0.3%이다.

본 연구에서 비교・분석하고자 하는 머신러닝 기반 예측 모델의 학습 및 테스트를 위하여 전체 데이터를 Table 1과 같이 분할하였다. 실제 운영 환경과 유사한 실험 환경을 구성하기 위하여, 2018학년도 봄학기부터 연속된 3개 학기의 데이터(학습 데이터 세트)로 머신러닝 기반의 예측 모델들의 구축을 위한 학습에 활용하고, 도출된 각 모델을 통해 2019학년도 가을학기 데이터(테스트 데이터 세트)를 기준으로 향후 1~3개 학기 이내에 중도 탈락할 학생을 예측하여 그 성능을 비교 분석한다.

Table 1. Data partitioning for training and test

data set	period	# of students (%)
training	the first semester of 2018~ the first semester of 2019	19,402 (75.9%)
test	the second semester of 2019	6,165 (24.1%)

한편 각각의 데이터 세트에 포함된 중도 탈락 학생 수 및 비율은 Table 2와 같다. 데이터 추출 시점을 기준으로 현재 재학 중인 학생이 1학기 이내에 이탈할 비율은 3.68%, 2학기 및 3학기 이내에 이탈할 비율은 각각 7.02%, 10.48%에 해당한다.

Table 2. Dropout students

data set	# of students who		
	drop within 1 semester	drop within 2 semesters	drop within 3 semesters
training	681 (3.51%)	1,412 (7.278%)	2018 (10.40%)
test	259 (4.20%)	430 (6.97%)	662 (10.74%)
total	940 (3.68%)	1,842 (7.20%)	2,680 (10.48%)

2.2 변수의 선정

본 연구에서는 A대학교 학사관리 시스템에서 추출할 수 있는 데이터 중 실험 기간 동안 데이터의 정합성과 연속성이 확보되는 25개의 독립변수와 예측 기간에 따른 3개의 종속변수를 선정하였다.

데이터 추출 시점을 기준으로 성별, 나이, 군필 여부, 생활관 거주 여부, 생활관 거주 기간, 재적 기간, 전과 여부, 학적 상태, 총 휴학 기간, 휴학 기간(데이터 추출 시점 휴학생의 경우), 학기별 평균 이수학점, 총 이수학점, 현 학기 전공 이수학점, 현 학기 교양 이수학점, 평균 성적, 현 학기 성적, 전공 평균 성적, 교양 평균 성적, 학기별 평균 상담 횟수, 현 학기 상담 횟수, 평균 출석률, 현 학기 출석률, 입학 장학금액, 평균 장학금액, 현 학기 장학금액 등 25개의 독립변수를 설정하였다. 한편 군필 여부와 같은 범주형 데이터(범주: 군필, 미필, 면제, 여학생 등)는 원-핫 인코딩(One Hot encoding)을 통해 변환하여, 모델에는 총 28개의 독립변수가 사용되었다.

종속변수로는 데이터 추출 시점을 기준으로 향후 1개 학기 이내에 중도 탈락 여부를 나타내는 DROPOUT_1과, 2개 또는 3개 학기 이내에 중도 탈락 여부를 나타내는 DROPOUT_2, DROPOUT_3을 도입하여, 각각의 종속변수와 28개의 독립변수로 구성된 모델을 구축하고 성능을 분석하였다. 종속변수를 시점에 따라 3개로 구분한 이유는 예측 모델이 단기(1개 학기 이내), 중기(2개 학기 이내) 및 장기(3개 학기 이내) 예측에 어느 정도 성능을 보이는지 확인해 보기 위함이다.

변수에 대한 설명은 Table 3에 정리되어 있다.

2.3 예측 모델의 구축

본 연구의 목적은 머신러닝 및 딥러닝 기반의 다양한 중도 탈락 예측 모델을 구축하고, 그 성능을 비교・분석하여 대학생 중도 탈락 예측에 가장 적합한 예측 모델을 제안하는 것이다. 이를 위해 머신러닝 분야에서 널리 활용되는 결정 트리(DT), 랜덤 포레스트(RF), 로지스틱 회귀(LR)와 딥러닝(DL)을 활용한 예측 모델을 구축하였다.

DT는 데이터의 학습을 통해 트리 구조의 분류 규칙을 생성하는 모델로 일반적으로 CART(classification and regression tree) 알고리즘을 사용하여 분류 규칙을 생성한다(Sang Bong Oh, 1996). RF는 복수의 결정 트리로 구성된 앙상블(ensemble) 방식의 모델로, 원본 데이터 세트에 대하여 중복을 허용하는 부트스트래핑(bootstrapping) 샘플 추출 방법을 적용하여 다수의 표본 집합을 구성하고 이를 기반으로 다수의 결정트리를 생성하여 예측에 활용하는 방식이다(Ohn et al., 2013). LR은 독립변수의 선형 결합을 이용하여 사건의 발생가능성을 예측하는 통계 기법이며, DL은 인공신경망의 층을 연속적으로 깊게 쌓

Table 3. Variables

variables	description
GENDER	male or female
AGE	age in the corresponding year
MILITARY_SERVICE	military service status (completion, unfinished, exemption, female student)
RESIDENTIAL_TYPE	whether the student lives in dormitory
DORMITORY_PERIOD	period (# of semesters) during the student has lived in the dormitory
ENROLLED_PERIOD	the student's enrollment period (# of semesters)
CHANGE_MAJOR	whether the student has changed one's major
STATUS	current academic status (in school or leave of absence)
NUM_ABSENCE	the # of semester during leave of absence
LATEST_ABSENCE	the last periods of leave of absence
AVG_CREDITS	average # of credits for each semester
TAKEN_CREDITS	total earned credits during enrollment period
COURSE_MAJOR	credits earned in major for the latest semester
COURSE_LIBERAL	credits earned in liberal art for the latest semester
AVG_GRADE	average grade for each semester
GRADE	grades for the latest semester
GRADE_MAJOR	major grades for the latest semester
GRADE_LIBERAL	liberal art grades for the latest semester
AVG_COUNSELING	average # of counselings with the adviser for each semester
LATEST_COUNSELING	# of counselings with the adviser for the latest semester
AVG_ATTENDANCE	average attendance rate for each semester
ATTENDANCE	average attendance rate for the latest semester
ENTERANCE_SCHOLARSHIP	admission scholarship amount
AVG_SCHOLARSHIP	average scholarship amount for each semester
SCHOLARSHIP	scholarship amount for the latest semester
DROPOUT_1	whether the student will drop out within following semester
DROPOUT_2	whether the student will drop out within following 2 semesters
DROPOUT_3	whether the student will drop out within following 3 semesters

아올려 데이터를 학습하는 모델로 최근 분류를 포함하여 이미지 인식 등의 다양한 분야에 폭넓게 활용되고 있다.

본 연구에서는 DT, RF, LR 기반의 모델의 경우 파이썬의 사이킷런(scikit-learn 1.2.2) 라이브러리를, DL 기반의 모델은 텐서플로우(tensorflow 2.10.0) 라이브러리를 사용하여 구현하였다.

한편 Table 2에서 보듯이 본 연구의 데이터 세트의 경우 중도 탈락 학생의 비율이 낮아(3~10% 내외) 종속변수의 범주(class) 간에 불균형한 분포를 갖는다. 머신러닝 분야에서 이러한 클래스 불균형 문제(class imbalance problem)를 해결하기 위한 일반적인 방법은 클래스 간 샘플 수의 비율을 맞추기 위한 샘플링 기법을 사용하는 것이다(Stjepan et. al., 2019). 본 연구에서는 SMOTE (Synthetic Minority Over-Sampling Technique)를 활용한 오버샘플링 기법을 활용하여 학습 데이터의 중도 탈락 샘플 수를 증식시켜 모델 구축에 사용하였다(Chawla et. al., 2002)

또한 구축된 각 모델의 성능을 평가하고 모델 간 성능을 비교 · 분석하기 위하여 머신러닝 분야의 보편적인 성능 평가 지표인 정확도(accuracy), 정밀도(precision), 재현율(recall) 및 f1-값(f1-value)을 사용하였다. 종속변수의 각 범주를 0(정상, 중도 탈락 안 함)과 1(중도 탈락)로 구분하고, 예측 모델의 예측 결과로 Table 4와 같은 오차행렬(confusion matrix)이 주어졌을 때 각 성능 지표의 값은 다음과 같이 계산된다.

Table 4. Confusion matrix

actual \ predicted	class 0: normal	class 1: dropout
class 0: normal	True Negative (TN)	False Positive (FP)
class 1: dropout	False Negative (FN)	True Positive (TP)

- 정확도 = $(TN+TP)/(TN+FP+FN+TP)$
- 정밀도 = $TP/(FP+TP)$
- 재현율 = $TP/(FN+TP)$
- f1-값 = (2×정밀도×재현율)/(정밀도+재현율)

더불어 분류 모델 간의 성능 비교로 많이 사용되는 ROC곡선(receiver operating characteristic curve)의 AUC (area under curve)도 함께 사용한다. AUC는 FPR(false positive rate)을 x축, TPR(true positive rate)을 y축으로 설정하고 모델의 성능을 ROC 곡선으로 나타내고 그 면적을 계산한 값이다.

한편, 본 연구에서처럼 클래스 불균형이 있는 데이터 세트에서는 정확도는 주요 지표로 간주되지 않는다 (Elrahman et al., 2013). 예측 모델이 다수의 샘플이 포함된 클래스(class 0)로만 예측하더라도 높은 정확도가 보장되기 때문이다. 본 연구와 같이 소수의 샘플이 있는 클래스(class 1)를 예측하는 것이 중요한 모델에서는 재현율이 중요한 성능지표가 될 수 있다. 중도 탈락할 학생을 음성(negative)으로 잘못 예측하게 되면 이들의 탈락을 예방할 기회를 잃게 되기 때문이다. 또한 일반적으로 재현율과 정밀도는 상쇄관계(trade-off)가 있기 때문에 이들을 조화 평균한 f1-값이나 AUC도 중요한 의미를 지닌다.

3. 성능 평가

본 연구에서는 결정트리, 랜덤 포레스트, 로지스틱 회귀 및 딥러닝 기반의 예측 모델을 구축하고 그 성능을 비교분석 하였다.

먼저 결정트리 기반의 예측 모델(DT-based Model)의 성능은 Table 5에 제시되어 있다. 데이터 추출 시점을 기준으로 장기 예측(3학기 이내에 중도 탈락 예측) 시 가장 높은 재현율과 f1-값 및 AUC를 보이고 있다. 반면, 1학기 이내의 단기 예측은 재현율과 정확도가 너무 낮아 현실적으로 활용에 한계가 있다.

Table 5. Performance of DT-based Model

	DROPOUT_1	DROPOUT_2	DROPOUT_3
accuracy	0.9186	0.9084	0.8874
precision	0.1093	0.3699	0.4778
recall	0.1313	0.4465	**0.5196**
f1-value	0.1193	0.4046	**0.4978**
AUC	0.5419	0.6947	**0.7259**

랜덤 포레스트 기반의 예측 모델(RF-based Model)의 성능은 Table 6에 제시되어 있다. 결정 트리 기반의 모델과 마찬가지로 장기(3학기 이내) 예측 시 가장 높은 재현율, f1-값 및 AUC 값을 보이고 있으며, 전반적인 성능은 결정트리 기반의 모델보다 우수한 것을 확인할 수 있다.

Table 6. Performance of RF-based Model

	DROPOUT_1	DROPOUT_2	DROPOUT_3
accuracy	0.9492	0.9401	0.9268
precision	0.2353	0.6027	0.7269
recall	0.0927	0.4163	**0.5106**
f1-value	0.133	0.4924	**0.5998**
AUC	0.7718	0.8782	**0.8807**

다음으로 로지스틱 회귀 기반의 예측 모델(LR-based Model)의 성능이 Table 7에 제시되어 있다. LR-based Model 구축 시 독립변수에 피처 스케일링(feature scaling) 기법을 적용하였다. 피쳐 스케일링 방법으로 표준화(standard normalization)와 정규화(min-max normalization)를 많이 사용하는데, Table 7에는 보다 높은 예측 성능을 보인 표준화 기법을 적용한 모델의 성능을 제시하였다. LR 기반 모델의 재현율의 경우 중기 예측 성능이 장기 예측보다 근소한 차이(0.02%p)로 높지만, f1-값은 장기 예측에서 훨씬 높게 나타나고 있다(8.39%p).

Table 7. Performance of LR-based Model

	DROPOUT_1	DROPOUT_2	DROPOUT_3
accuracy	0.8394	0.8209	0.8071
precision	0.1434	0.2278	0.3111
recall	0.5676	**0.6558**	0.6556
f1-value	0.229	0.3381	**0.422**
AUC	0.7826	**0.8396**	0.8352

마지막으로 Table 8에 딥러닝 기반의 중도 탈락 예측 모델(DL-based Model)의 성능평가 결과를 제시하였다.

Table 8. Performance of DL-based Model

	DROPOUT_1	DROPOUT_2	DROPOUT_3
accuracy	0.9421	0.9377	0.9226
precision	0.1938	0.5697	0.7051
recall	0.1197	0.4372	**0.4804**
f1-value	0.148	0.4947	**0.5714**
AUC	0.7427	**0.8485**	0.8304

본 연구에서 구축한 딥러닝 기반의 예측 모델의 구조는 Fig. 1에 도식화되어 있다. 활성함수로 중간 layer에서는 relu를, 마지막 layer에서는 sigmoid 함수를 사용하였고, 독립변수를 대상으로 피처 스케일링 기법 중 정규화를 수행하였다. 참고로 실제 실험에서는 은닉층의 개수와 층별 노드의 개수를 달리하여 다양한 구조의 모델을 구축하고 성능을 비교하였지만 큰 차이가 나타나지는 않았다.

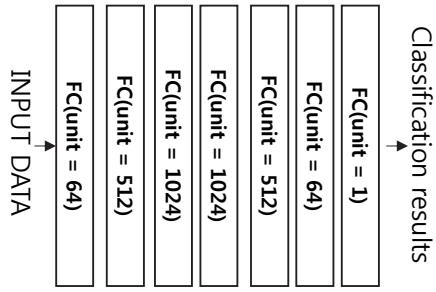

Fig. 1. Architecture of DL-based Model

Table 8에서 보듯이 DL-based Model에서도 3학기 이내에 중도 탈락(DROPOUT_3)을 예측하는 모델이 가장 높은 재현율 및 f1-값을 보이고 있다.

한편 앞서 실험한 모든 모델들은 전반적으로 예측 기간이 길어질수록 좋은 예측 성능을 보인다. 이는 예측 모델을 통해 중도 탈락 가능성이 크다고 식별된 학생들이 종종 즉시 이탈하는 것이 아니라 점진적으로 이탈하는 경향을 보인다는 사실로 설명될 수 있다. 또한 대학 측면에서는 이러한 학생들을 적극적으로 관리함으로써 중도 탈락을 예방할 시간적 기회를 확보할 수 있음을 의미한다.

앞서 구축한 모델들의 성능을 비교한 결과가 Fig 2.에 제시되어 있다. Fig 2.의 결과는 DROPOUT_3을 종속변수로 할 때 각 모델의 성능을 비교한 것이다.

Fig. 2에서 보듯이 장기 예측에 있어, 랜덤 포레스트 기반 모델은 재현율을 제외한 모든 성능 지표의 값이 가장 높지만, 재현율은 3순위로 그다지 높지 못하다. 반면 타 모델에 비해 재현율이 월등히 높은 로지스틱 회귀 기반 모델은 정밀도 및 f-1값이 가장 떨어짐을 알 수 있다. 이는 로지스틱 회귀 모델이 오버샘플링으로 인해 실제 원본 데이터의 유형보다 훨씬 많은 중도 탈락 데이터를 학습하면서, 테스트 데이터 세트에서 정밀도가 현저히 떨어진 것으로 해석할 수 있다. 한편 최근 널리 활용되는 딥러닝 기반의 예측 모델은 중도 탈락 예측에서는 그다지 높은 성능을 보이지 않았다. 이는 최근 kaggle(https://www.kaggle.com) 등 머신러닝 대회의 정형화된 데이터의 분류 분석에서 앙상블 기반의 모델들이 딥러닝 기반 모델보다 높은 성능을 보이고 있다는 사실과 일맥상통한다.

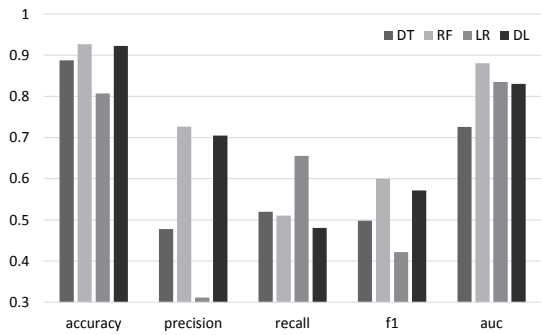

Fig. 2. Comparison of prediction models

본 연구는 중도 탈락 가능성이 높은 학생들을 조기에 파악하고, 이들을 사전 관리하여 중도 탈락 비율을 줄이는 것을 목적으로 한다. 따라서 각 모델의 성능 비교는 재현율을 기반으로 판단하는 것이 적절하다고 할 수 있다. 그러나 재현율만 높고 정밀도가 낮은 경우 예측 모델은 상당수의 이탈 후보군을 생성하게 되고, 이는 해당 학생들의 사전 관리에 많은 자원이 투입되어야 함을 의미한다. 따라서 각 대학이 보유한 상담 인력 등의 자원을 고려하여, 투입되는 자원의 양과 상관없이 철저한 사전 관리를 수행하고자 한다면 로지스틱 회귀 기반의 예측 모델이 적합할 것이고, 보유 자원 대비 효율적인 관리를 하고자 한다면 랜덤 포레스트 기반 모델을 사용하는 것이 바람직하다고 하겠다.

4. 결론

본 연구에서는 설문조사를 기반으로 중도 탈락의 원인을 규명하고자 했던 다수의 기존 연구와는 달리, 대학이 보유한 학사관리 시스템의 정량적 자료를 기반으로 실제 활용 가능한 중도 탈락 예측 모델을 제안하였다. 이를 위해 머신러닝 분야에서 폭넓게 활용되는 결정 트리, 랜덤 포레스트, 로지스틱 회귀 기반의 모델과 최근 널리 적용되는 딥러닝 기반의 모델을 구축하고 그 성능을 비교·분석하였다.

분석 결과 로지스틱 회귀 기반의 예측 모델이 재현율 측면에서 가장 뛰어난 성능을 보였으며, 재현율을 제외한 나머지 모든 지표에서 랜덤 포레스트 기반 모델이 가장 우수했다. 또한 종속변수를 예측 기간에 따라 3가지로 설정하고 분석한 결과, 모든 모델에서 장기 예측 시 가능 높은 성능을 보였다.

본 연구의 결과를 바탕으로 대학들은 각자의 실정에 맞는 예측 모델을 구축하여 중도 탈락이 예상되는 학생들을 선별하고 관리한다면, 중도 탈락 비율을 줄이고 나아가 재정적 문제를 완화하여 지속 가능한 운영을 추구하는데 도움이 될 것으로 기대된다.

한편, 이 연구에서는 몇 가지 제한 사항이 존재한다. 우선, 본 연구의 결과는 A대학교의 데이터에 국한되어 있어 다양한 대학들의 상황을 대표하기에는 한계가 있다. 또한 데이터의 수집과정에서 발생한 제약으로 인해 일부 중요한 변수나 기간을 고려하지 못한 점도 한계로 지적된다. 더불어 최근에 발전된 앙상블 기반의 모델들과 새로운 머신러닝 기법에 대한 평가는 본 연구에서 다루지 못한 부분이다.

또한, 본 연구에서 제안된 예측 모델들의 정확도는 80-94% 정도로 높은 편이지만, 재현율은 모델별로 차이는 있지만 최대 65.56% 정도에 그치고 있다. 이는 학적 데이터만으로는 중도 탈락 예측의 성능을 월등히 높이기에는 한계가 있음을 의미한다고 하겠다. 최근 대학에서는 학생들과의 상담이 강조되고 있는데, 이러한 종류의 추가 데이터를 함께 분석한다면 보다 높은 예측 성능을 달성할 수 있을 것으로 기대된다.

향후 더 많은 데이터 수집과 다양한 최신 기법들을 활용한 모델들의 분석을 통해 중도 탈락 예측의 성능과 신뢰성을 높이는 방향으로 노력할 계획이다.

References

Andrade, M. S., Miller, R. M., McArthur, D., & Ogden, M., “The impact of learning on student persistence in higher education”, Journal of College Student Retention: Research, Theory & Practice, Vol. 24, No. 2., pp. 316-336. 2020.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP., “SMOTE: Synthetic minority over-sampling technique”, J Artif Intell Res, Vol. 16, pp. 321-57, 2002.

Chung J. Y, M. S. Sun, and M. J. Jeong, “An Analysis of Institutional Factors Affecting on College Dropout Rates”, Asian Journal of Education, vol. 16, no. 4, pp. 57-76, 2015.

Elrahman S.M.A. and A. Abraham, “A Review of Class Imbal-ance Problem”, Journal of Network and Innovative Computing, Vol. 1, pp. 332-340, 2013.

Han S., “Exploration of Factors that Affect College Student Drop-out and Resilience,” Journal of Learner-Centered Curriculum and Instruction, Vol. 18, No. 24, pp. 1369-1391, 2018.

Stjepan Picek, Annelie Heuser, Alan Jovic, Shivam Bhasin, and Francesco Regazzoni. “The curse of class imbalance and conflicting metrics with machine learning for side-channel evaluations”, IACR Transactions on Cryptographic Hardware and Embedded Systems, Vol. 1, pp. 209-237. 2019.

Jeong, Do-Heon and Ju-Yeon Park, “Data Analysis of Dropouts of University Students Using Topic Modeling”, Jornal of the Korea Institute of Information and Communication Engineering, Vol. 25, No. 1, pp.88-95, 2021.

Jeong Do-Heon, “Implementation of a Machine Learning-based Recommender System for Preventing the University Students’ Dropout”, Journal of the Korea Convergence Society, Vol. 12, No. 10, pp.37-43, 2021.

Jeong, Seon-Ho. “A Study on the Development of University Students Dropout Prediction Model Using Classification Technique.” Journal of Convergence Consilience. Korea Safety Culture Institute, August

31, 2022.

Kang M., E. Lee, and E. Lee, “Trends and influencing factors of college student’s dropout intention”, In Forum for Youth Culture, no. 58, pp. 5-30, 2019.

Park C., “Development of Prediction Model to Improve Dropout of Cyber University”, Journal of the Korea Academia-Industrial Cooperation Society, vol. 21, no. 7, pp. 380-390, 2020.

Lee E. H. and S. Kang, “The Research Trends and Implications of College Dropouts in Korea”, Journal of Learner-Centered Curriculum and Instruction, Vol. 19, No. 10, pp. 169-199, 2019.

Lee E., Y. Song, J. Kim, and S. Oh, “An Exploratory Study on Determinants Predicting the Dropout Rate of 4-year Universities Using Random Forest: Focusing on the Institutional Level Factors”, Journal of Educational Technology, Vol. 36, No. 1, pp. 191-219, 2020.

Lee S. and L. Park, “Analysis of Correlation between the Characteristics of University Students and Dropout”, Journal of Learner-Centered Curriculum and Instruction, Vol. 19, No. 11, pp. 1185-1210, 2019.

Ohn Syng-Yup, Seung-Do, Chi and Mi-Young Han, “Feature Selection for Classification of Mass Spectrometric Proteomic Data Using Random Forest” Journal of the Korea Society for Simulation, Vol. 22, No. 4, pp.139-147, 2013

Sang Bong Oh and Kun Chang Lee, “A Neural Network-Driven Decision Tree Classifier Approach to Time Series Identification”, Journal of the Korea Society for Simulation, Vol. 5, No. 1, pp.1-12. 1996.

정 석 봉 (ORCID : https://orcid.org/0000-0002-6209-1935 / sbjung@kiu.ac.kr)

1999 한국과학기술원(KAIST) 산업경영학과 학사
2001 한국과학기술원(KAIST) 산업공학과 석사
2005 한구과학기술원(KAIST) 산업공학과 박사
2011~ 현재 경일대학교 철도학부 교수

관심분야 : 이미지 인식, 딥러닝, 컴퓨터 비전, 사회네트워크분석

김 두 연 (ORCID : https://orcid.org/0000-0001-8750-4444) / duyonkim@kiu.ac.kr)

2003 연세대학교 사회환경시스템공학부 공학사
2005 연세대학교 토목환경공학과 공학석사
2009 연세대학교 토목환경공학과 공학박사
2011~ 현재 경일대학교 건축토목공학과 교수

관심분야 : 위험관리, 해외건설, 의사결정지원시스템, 레질리언스