산업융합연구 제22권 제2호 pp. 23-32, 2024
ISSN 2635-8875 / e-ISSN 2672-0124
DOI : https://doi.org/10.22678/JIC.2024.22.2.023

전문대학 학생의 학업중단 예측에 관한 연구: 초기 학업 성적의 중요성

오상조[1], 심지환[2*]
[1]동양미래대학교 경영정보학과 교수, [2]동양미래대학교 빅데이터경영과 조교수

A Study on Predicting Student Dropout in College: The Importance of Early Academic Performance

Sangjo Oh[1], JiHwan Sim[2*]
[1]Professor, Dept. of Management Information Systems, Dongyang Mirae University
[2]Assistant Professor, Dept. of Big Data Management, Dongyang Mirae University

요 약 본 연구에서는 서울 소재 한 전문대학 학생들을 대상으로 하여 최소한의 인구통계학적 변수와 1학년 1학기 성적을 활용하여 학생들의 최종 학적 상태를 예측하고자 하였다. XGBoost와 LightGBM 모델을 사용한 결과, 이러한 변수들이 학생들의 제적 여부 예측에 유의미한 것을 발견하였다. 이는 학업 시작 초기의 성적이 학업 중단의 중요한 지표가 될 수 있음을 시사한다. 또한, 전문대학의 학제가 최종 학적에 영향을 미칠 가능성을 확인하였으며, 이는 학업 기간이 학생들의 학업 중단 결정에 중요한 요소임을 나타낸다. 전문대학에서 조기 학업 중단 의도를 파악하는 데 있어 심리적, 사회적, 경제적 요인에 의존하지 않고 학업 성취도만을 기준으로 모델링을 시도하였다. 이는 향후 학업 중단에 대한 조기 경보 시스템 구축에 도움이 될 것으로 기대된다.

키워드 : 전문대학, 학업중단, 조기예측, Random forest, XGBoost, LightGBM

Abstract This study utilized minimum number of demographic variables and first-semester GPA of students to predict the final academic status of students at a vocational college in Seoul. The results from XGBoost and LightGBM models revealed that these variables significantly impacted the prediction of students' dismissal. This suggests that early academic performance could be an important indicator of potential academic dropout. Additionally, the possibility that academic years required to award an associate degree at the vocational college could influence the final academic status was confirmed, indicating that the duration of study is a crucial factor in students' decisions to discontinue their studies. The study attempted to model without relying on psychological, social, or economic factors, focusing solely on academic achievement. This is expected to aid in the development of an early warning system for preventing academic dropout in the future.

Key Words : College, Academic dropout, Early prediction, Random forest, XGBoost, LightGBM

This paper was supported by Dongyang Mirae University research fund in 2023.
*Corresponding Author : JiHwan Sim(sim2080@dongyang.ac.kr)
Received December 28, 2023 Revised January 10, 2024
Accepted February 20, 2024 Published February 28, 2024

1. 서론

최근 통계청에 따르면, 한국은 2017년을 기점으로 합계출산율이 1명 미만으로 떨어지고, 2022년에는 약 0.78명을 기록하는 초저출산 시대에 돌입하게 되었다. 통계청에 따르면, 2021년도 한국의 합계출산율은 약 0.81명이었는데, 이 수치는 OECD 평균의 절반을 간신히 기록한 수치이기도 하였다[28]. 더불어 교육부에 의하면, 전문대학의 재학생 충원율은 2019년 110.2%를 기록하고, 이후 감소 추세에 들어서 2022년에는 103.3%를 기록하였고, 학업 중단율은 2011년 7.1%를 기록한 이래로 꾸준히 증가하여 2022년에는 8.1%를 기록하였다. 이러한 추세는 비단 전문대학에서만 나타는 현상이 아닌, 일반대학, 교육대학 등 여러 고등교육기관에서 유사하게 나타나고 있다. 하지만 고등교육기관으로의 진학 포기 및 진학 이후의 이탈 문제는 전문대학에서 더욱 두드러지게 나타나고 있는 것이 현재 한국 고등교육기관의 현실이다.

전문대학의 교육과정은 직업기술 지향적 교육에 초점을 두어 현장 중심 교육 및 실무 중심의 교육이 크게 활성화 되어있다. 그러한 이유에서 실무기술을 지닌 노동인구를 노동시장에 빠르게 공급해 주는 전문대학의 역할은 노동시장에서 요구하는 다양한 노동수요를 충족시켜 주는데 큰 역할을 하고 있다. 그럼에도 불구하고 전문대학의 재학생 충원율 및 학업 중단율은 일반대학에 비해 상대적으로 빠르게 악화하고 있어 당해 대학들의 존재 가치 및 역할이 자연스럽게 축소될 상황에 직면해 있다. 이는 대학의 재정적 안정성에도 큰 위협으로 작용하고, 그로 인한 교육 서비스 질의 저하로 이어지는 악순환이 발생할 수 있다. 우리나라 전문대학은 대부분 사립이며 등록금 의존도가 매우 높아 재학생 유지율은 대학의 생존과 직결될 수 있는 문제이기도 하다.

그리고 학업 중도 이탈에 대한 경각심이 확대되는 상황에 여러 선행연구에서 학업 중단은 대개 1학년 재학 중에 많이 발생하고 있으며, 학업 중단 요인으로 다양한 심리적, 사회적, 경제적 요인들이 복잡하게 얽혀 작용하고 있음을 밝혀냈다[5, 25]. 그리고 이러한 결과는 입학부터 졸업까지의 학업 유지 기간이 길지 않은 전문대학에서는 학생의 학업 중단 및 이탈을 더욱더 조기에 포착해야 할 중요한 이유가 생긴 셈이다.

이러한 맥락에서 본 연구는 서울에 소재한 한 전문대학의 입학생 학적 자료를 토대로 학업 중단 가능성이 있는 학생들을 조기에 포착하기 위한 중요변수를 식별하고, 예측분석을 수행하였다. 본 연구는 서울 소재 A 전문대학 학생들의 학적 자료를 토대로 세 가지 기계학습 알고리즘으로 두 가지 학업 중단 예측 모델을 실험하였다.

논문의 구성은 다음과 같다. 먼저, 학업 중단 요인 식별과 학업 중단 예측과 관련된 문헌들을 검토한다. 그 후, A 전문대학이 보유한 10년간의 학적 데이터의 수집, 정제, 분석 과정에 대해 상세히 기술하고 마지막으로 분석 결과 및 본 연구의 의의와 한계에 대해서 논의하도록 한다.

2. 선행연구

2.1 학업 중단 요인 식별에 관한 연구

학업 중단은 대학에서 학생이 자신의 학위 과정을 완료하지 않고 중단하는 것을 의미하며, 등록 시에 약술된 사전 결정된 교육목표에 대한 특정 역량의 개발과 관련하여 개인의 실패를 의미한다[24]. 이는 학생이 대학 교육과정에서 체계적인 학습에 참여하지 않는 상황으로, 여러 복합적인 요인에 의해 발생할 수 있으며, 단순히 학교를 그만두는 것 이상의 의미를 가진다. 이처럼 학업 중단에 대한 정확한 개념적 정의를 내리는 것은 이론적 영역을 넘어서 전 세계 대학과 국가에서 개발한 정책, 행동, 연구에서 나타나는 복잡한 작업이다[13]. 학업 중단 현상을 설명하기 위해 심리학, 사회학, 경제학 등 최소한 세 가지 분야에서 광범위하게 연구되고 있다[2]. 그리고 각 분야에서 활용되는 요인들이 상호작용하는 형태로써 학업 중단을 설명하는 데 활용될 수 있다[8, 26].

학업 중단을 개인의 심리적 성향에 초점을 둔 연구는 아주 오래전부터 진행됐다. 학위를 마치기 전에 대학을 중퇴하는 학생들의 경향을 4년간 추적하여 대상 학생들의 특징을 조사한 결과, 중퇴자들의 심리적 성향은 비중퇴자보다 더 냉담하고, 자기중심적이며, 충동적이고 자기주장이 강한 경향이 있다고 보고하였다[5]. 대인관계 특성인 친절함, 성실함, 정서적 안정성, 외향성, 개방성과 더불어 공격성, 진로 결정성, 낙관주의, 자기주도학습, 정체성 의식, 강인한 마음가짐, 일에 대한 동기와 같은 좁은 성격 특성을 검토하였다[14]. 그리고 학업을 중단하려는 의도를 가진 학생들과 정체성 의식과 정서적 안정성, 일에 대한 동기가 유의미한 관련성이 있음을 발견하였다.

대학교 입학 후의 생활은 학생들에게 중요한 사회적 경험을 제공하게 된다. 이 시기에 학생들은 정서적 안정

성, 직업에 대한 동기, 진로 결정과 같은 심리적 요소뿐만 아니라 가정환경, 부모님의 지지와 같은 사회적 요인, 교육기관의 환경과 장학금 수여 여부와 같은 경제적 요인의 영향을 받는다. 이러한 요인들은 학생들의 학업 성취와 밀접하게 연관되어 있으며, 그 영향을 개별 학생의 상황에 따라 다를 수 있다. 가령, 학생 자신이 선택하려는 학습 프로그램 또는 전공의 특성에 대해 완벽하게 파악하지 못한 상황에서 학업을 완수하는 데까지 필요한 노력이나 교우관계, 예상 비용 등을 올바르게 평가하지 못할 수 있다[2]. 이러한 사회, 경제적 불확실성 속에서 학생의 중퇴 의도는 강해질 수 있고, 이는 자기주도 학습 능력과 자기 효능감과 같은 학업적 요소들에 부정적 영향을 줄 수 있다[27].

국내 대학에 재학중인 학생들은 당해 학교의 모든 교육프로그램을 이수하는 동안 여러 학기를 거치며, 중간/기말 단위로 평가된다. 그리고 학생들의 학업 이수 과정 중 최종 학업 및 학적 상태 또는 첫 학기 이외의 학기에서 이전 학기의 자료를 토대로 다음을 예측할 수 있는 기회를 가질 수 있다[4]. 학생의 성과는 사회경제적, 심리적, 환경적 요인의 산물이라는 점에서 학생의 중퇴 의도를 파악하는데 매우 중요하게 활용된다[17]. 학생의 성과는 학생의 발전을 의미하며, 이를 식별하기 위해 다양한 학업 요인을 고려할 수 있는데, 미국의 경우 대표적으로 GPA를 가장 많이 활용하며, 국내의 경우에 대응되는 요소가 수능 성적이라 할 수 있다. 이 외에도 교육과정에 대한 참여정도(출결), 대학에서의 분기별 성적(중간/기말) 및 학기별 성적, 최고 시험점수 등이 활용된다[3].

2.2 학업 중단 예측에 관한 연구

OECD 회원국의 최근 통계에 따르면 대학에 입학하는 학생 중 1/3이 학위를 취득하지 못한 채 학업을 중단하는 것으로 나타났다[19]. 많은 미주, 유럽 국가들에서도 2학년이 시작될 무렵에 중도 탈락하게 되는 비율이 평균적으로 10%를 웃돌고 있다[2]. 학생들의 중도 탈락은 국가별로 교육 체계, 사회구조, 경제 상황 등 서로 다른 원인이 있을 것이다. 하지만 대학의 생존이 달린 학생 유지율의 관점에서 학생의 이탈을 방지하고자 다양한 방식으로 이탈 가능성이 높은 학생들을 사전에 파악하고, 다시 학교로 유인할 방법을 고민하는 것은 모든 학교에서의 공통 관심사로 대두되고 있다. 그리고 고등교육기관이 보유한 막대한 양의 과거 및 현재 학생 데이터를 토대로 학생의 성공에 영향을 미치는 특징을 식별하는 데 큰 노력을 기울이고 있다[1, 11].

학생 유지에 관한 연구는 전통적으로 설문조사를 중심으로 이루어졌다. 예를 들어, 학생 집단을 대상으로 설문조사를 실시하고 특정 기간 이들을 추적하여 교육을 계속 이어나갈지 여부를 응답받는 형태로 이루어진다. 이러한 설계를 토대로 연구자들은 이론적 모델을 개발하고 검증하는 작업을 진행해 왔다. 즉, 이론적 연구는 데이터 기반의 분석 연구에 사용할 중요한 예측 변수를 식별하는 데 도움이 될 수 있다[10]. 다양한 연구들에 의해 기반이 마련된 이론적 틀하에서 교육환경에 적용될 수 있는 새로운 연구분야로 교육데이터마이닝(EDM, Educational Data Mining)이 떠오르고 있다. 이는 학생 데이터 행정 기록 등을 포함하여 기관에서 얻은 방대한 양의 데이터를 사용하며, 교육기관에서 학생의 행동을 이해하고 학업 성공 및 실패 학생을 식별하여 교수 및 학습환경을 개선하는 데 활용될 수 있다[21].

최근 몇 년 동안 대규모 온라인 공개과정(MOOC) 및 기타 온라인 교육 환경의 증가로 인해 EDM 및 학습분석 영역에서 교육 데이터에 대한 기계학습 기술 적용 사례가 증가했다[6]. 학생 중퇴 예측을 다루는 교육의 맥락에서의 기계학습 기술은 크게 지도학습(Supervised Learning)과 비지도학습(Unsupervised Learning)으로 나눌 수 있다. 지도학습은 가능한 가장 높은 정확도로 실험 데이터에서 레이블이 지정되지 않은 예제를 식별할 수 있도록 훈련 데이터를 학습하는 것을 기반으로 한다[16]. 반면, 비지도학습은 레이블이 지정되어 있지 않은 데이터를 토대로 유사한 성격을 가지는 관측치들을 군집화하거나 각 관측치 간 관계를 식별하는 방법이다. 이러한 기술이 강조되는 이유는 학생들이 미래를 계획하고 학업 경력을 향상하는 데 도움이 될 수 있는 예측 시스템을 구축할 가능성이 있기 때문이다[15].

EDM 분야에서 빈번히 연구되는 주제 중 하나는 학생의 학업 성공 또는 실패를 예측하는 것이다. EDM은 정량 데이터 기반의 방법론으로써 학생의 학업 성공 여부를 예측할 만한 신뢰성이 확보되면 이를 토대로 교육과정을 수정하거나 학업 성과가 약한 학생에게 피드백을 제공하는 것이 가능하다[18]. 즉, EDM의 도구로 활발히 활용되는 기계학습은 학생 중퇴에 대한 예측 모델을 구축하기 위한 유망한 도구이며, 중퇴 위험이 있는 학생들에게 조처하도록 관련 기관에 조기 경고를 제공할 수 있다[9]. 특히, 데이터 확보가 용이한 e-learning에 참여한 학생의 학업성

과지표를 대상으로 대표적인 머신러닝 기법인 Decision Tree(DT), Support Vector Machine(SVM), Naive Bayes(NB) 등의 알고리즘을 적용하여 75% 이상의 중퇴 예측율을 보인 많은 연구가 존재한다[15, 22].

3. 데이터 및 연구방법

3.1 데이터

본 연구의 원본 데이터는 2010년부터 2023년까지의 입학생 전체 37,917명에 대한 서울 소재 A 전문대학이 보유한 2023년 기준의 학적자료이다. 총 275개의 변수가 입력되어 있으며, 각 변수는 인적정보, 출신고교, 학과, 학기별 이수과목에 대한 과목수와 학점, 평점평균, 강의평가, 휴학기간, 현재학적 정보 등으로 구성되어 있다. 본 연구는 중퇴를 조기 예측 가능성을 검증하기 위해 학생의 입학 후 가장 먼저 부여받는 학업 성과 지표인 1학년 1학기 성적을 대상으로 실험 모델을 구성하였다. 실험은 세 개의 기계학습 알고리즘을 토대로 변수의 조합을 달리 구성하여 두 가지 모델을 실험하며, 각 모델에 활용되는 변수와 예측 변수는 Table 1에 정리되어 있다.

3.2 연구설계

Fig. 1은 23년도 기준 최종학적별 학생 수를 보여주고 있고, 19년도부터 23년도까지의 입학생의 최종학적과 이전 연도 입학생의 최종학적의 양상이 다름을 알 수 있다. 이는 학생들의 군휴학, 질병휴학, 가사휴학 등 불가피한 개인 사정으로 졸업 또는 제적이 유보된 결과로 해석할 수 있다. 따라서, 제적생 혹은 졸업생의 효과적인 예측을 위해 최종학적 간 특성이 명확히 드러나는 기간인 2010년부터 2018년까지의 자료를 훈련용 데이터셋으로 사용하였다. 그리고 2018년 이후의 기간 중 휴학생과 재학생의 비율이 높은 2022년, 2023년을 제외한 나머지 기간인 2019년부터 2021년까지의 자료를 실험용 데이터셋으로 사용하였다.

국가 및 교육기관과 같은 환경적 요인에 따라 학적 자료 기반의 기계학습 성능은 다수의 연구에서 다소 상이한 결과를 보인다. 또한 기계학습 알고리즘을 활용한 예측 모델 개발에서 가장 큰 장애물은 활용 데이터를 불문하고 결과에 대한 해석이 어렵고, 접근 방식이 블랙박스라는 점에서 비판을 받는다[20]. 그러한 이유에서 각 알고리즘이 예측한 학적 상태에 대한 근거 변수를 확인할 수 있고, 보편적으로 우수한 성능을 보인 Random Forest(이하 RF), XGBoost(이하 XGB), LightGBM(이하 LGBM) 세가지 알고리즘을 토대로 본 연구의 실험을 설계하였다.

본 연구는 학생들의 인구통계적 자료와 학업 초기의 평점정보가 학적예측에 유의미한 영향이 있을 것이라는 근거하에 시작되었다. 따라서, Table 1에 제시된 바와 같이 D, E1, E2, C1, C2, S를 모두 활용한 모델(이하 M1)과 D, S만을 활용한 모델(M2) 두 가지로 구성하였다. 각 모델을 위와 같이 구성한 이유는 각 모델에서의 성능에서 우려될 정도의 차이가 없다면, 이는 M1과 M2에서 공통적으로 활용된 변수가 더욱 중요하다는 것으로 해석할 수 있기 때문이다.

한편, 모든 실험에 대한 일반화된 성능을 확인하기 위

Table 1. Model-specific utilization variable

Variable Type	Variable Name	M1	M2
Demographic Data (D)	Gender	○	○
	High School Affiliation	○	○
	School System	○	○
Number of Subjects Enrolled (E1)	Major(Required/Select) Elective (Required/Select)	○	
Credit of Subjects Enrolled (E2)	Major(Required/Select) Elective (Required/Select)	○	
Number of Subjects Completed (C1)	Major(Required/Select) Elective (Required/Select)	○	
Credit of Subjects Completed (C2)	Major(Required/Select) Elective Required/Select	○	
Score (S)	Major Average(Requried/Select) Elective Average(Requried/Select) Total Major Average Total Elective Average Total Average	○	○
Academic Status (A)	Current Academic Status	●	●

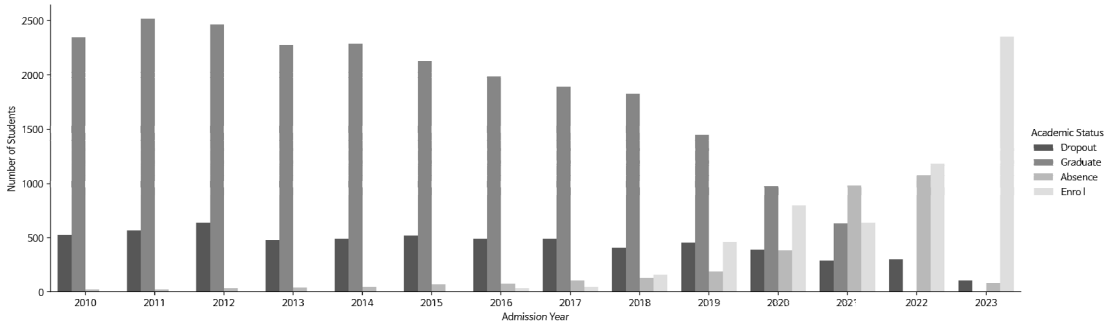

Fig. 1. Number of students about academic status by year

해서는 훈련용 데이터셋의 일부를 검증용 데이터셋으로 구분하여 성능을 모니터링해야 한다. 이를 위해 교차검증을 수행해야 하는데, 2010년부터 2018년까지의 자료에서 졸업생과 제적생의 비율이 매우 불균형한 모습을 보인다. 따라서, 각 모델에 대해 예측해야 할 범주의 비율을 고려한 Stratified K-Fold 샘플링 방식을 10번 적용하여 교차검증을 수행한다. 각 모델의 성능에 대한 지표는 Accuracy, Precision(정밀도), Recall(재현율), F1-score 네 가지 성능지표를 사용하였다.

본 연구에서는 Recall(재현율)과 Precision(정밀도)을 주목해야 한다. 먼저, Recall은 실제 제적 혹은 졸업생을 올바르게 식별하는 능력을 측정한다. 가령, 제적 사례를 놓치는 것이 중대한 영향을 끼치는 상황에서 중요한 의미를 지니며, 이는 가능한 많은 실제 제적 사례를 포착하는 것을 목표로 한다. 반면, Precision은 모델이 제적생을 예측한 경우 해당 사례가 실제 제적생인 비율을 평가한다. 이는 모델의 예측이 얼마나 정확한지를 보여주며, 잘못된 식별을 최소화하는 데 중요한 지표로 활용될 수 있다.

더불어, 각 모델에서 중요하게 활용된 변수를 식별하기 위하여 SHAP(SHapley Additive exPlanantions) value를 검토한다. SHAP value는 기계학습 모델의 예측에 대한 각 특성의 기여도를 설명하는 데 사용되는 개념으로써, 제적 확률에 대한 특성들의 영향의 정도와 분포를 확인할 수 있는 장점이 있다. 위 내용을 토대로 본 연구의 연구모형을 도식화하면 Fig. 2와 같다.

4. 연구결과

본 연구의 실험은 Python 프로그래밍 언어(버전 3.10.12)로 기계학습 프레임워크인 PyCaret(버전 3.2.0)과 Scikit-Learn(버전 1.2.2)을 활용하였다. 각 프레임워크는 구글에서 제공하는 클라우드 기반의 기계학습 작업

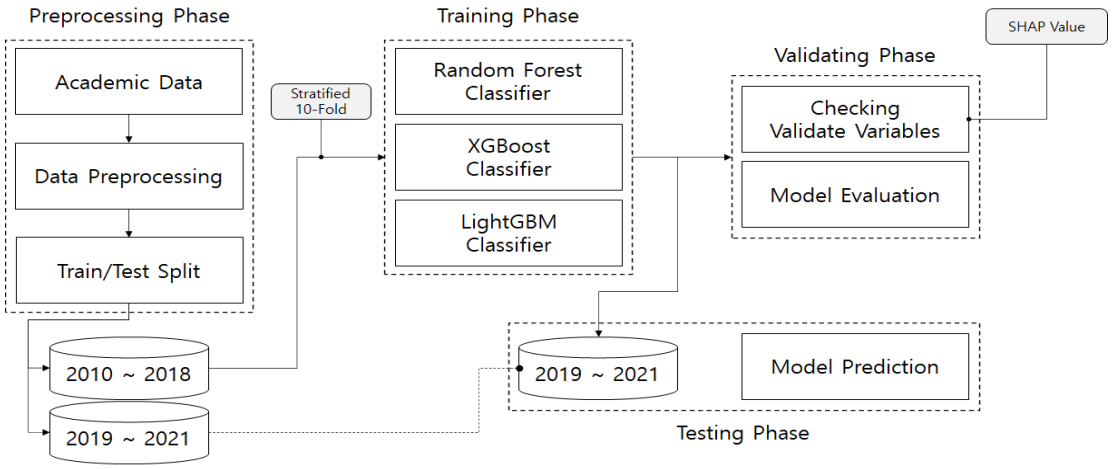

Fig. 2. Research process

환경인 Colaboratory(Colab)에서 수행되었다.

4.1 학업중단 예측 모델에 대한 성능평가

Table 2는 M1과 M2에서 활용된 기계학습 알고리즘별 제적과 졸업에 대한 실제 예측결과를 혼동행렬로 표현하고 있다. Table 2에 제시된 바와 같이, 분류에 성공한 비율을 제적과 졸업에 대해 모두 고려해 보면, M1과 M2에서 모두 RF에 비해 상대적으로 XGB, LGBM이 우수한 성능을 보인다.

Table 2. Confusion matrix of each models

		M1		M2	
RF	Dropout	264	408	277	395
	Graduate	100	3519	197	3422
XGB	Dropout	280	392	261	411
	Graduate	110	3509	67	3552
LGBM	Dropout	260	412	267	405
	Graduate	70	3549	88	3531
		Dropout	Graduate	Dropout	Graduate

Table 3. Evaluation matrix of each models

		Precision		Recall	
		Dropout	Graduate	Dropout	Graduate
M1	RF	0.725	0.896	0.393	0.972
	XGB	0.718	0.900	0.417	0.970
	LGBM	0.788	0.896	0.387	0.981
M2	RF	0.584	0.897	0.412	0.946
	XGB	0.752	0.897	0.397	0.976
	LGBM	0.796	0.896	0.388	0.981

이를 면밀히 살펴보기 위해 각 혼동행렬에 대한 Precision과 Recall을 계산하여 Table 3에 기술하였다.

먼저, Table 3에서 제적(Dropout)에 대한 성능을 살펴보면, M1에서의 Precision은 LGBM이 0.788, Recall은 XGB가 0.417을 기록하고 있다. 한편, M2에서의 Precision은 일관되게 LGBM이 0.796으로 비교적 우수한 성능을 보이고 있으나, Recall은 RF가 0.412를 기록하여 이외 알고리즘에 비해 비교적 높은 성능을 보였다. 이어 졸업(Graduate)에 대한 성능을 확인해보면 M1에서의 Precision은 XGB가 0.900, Recall은 LGBM이 0.981로 각각 가장 우수한 성능을 보였다. 한편, M2에서의 Precision은 RF와 XGB가 0.897로 거의 유사한 성능을 보였고, Recall은 LGBM이 0.981로 일관되게 가장 우수한 성능을 보였다.

위 성능평가 결과를 종합하면 대체로 XGB와 LGBM가 비교적 우수한 성능을 보이고 있음을 알 수 있다. 따라서, 각 모델에서의 XGB와 LGBM에 유의미하게 활용된 변수들을 확인하여 학업중단 예측을 위한 중요 변수가 무엇인지 식별하고자 하였다.

4.2 학업중단 예측에 대한 중요변수의 식별

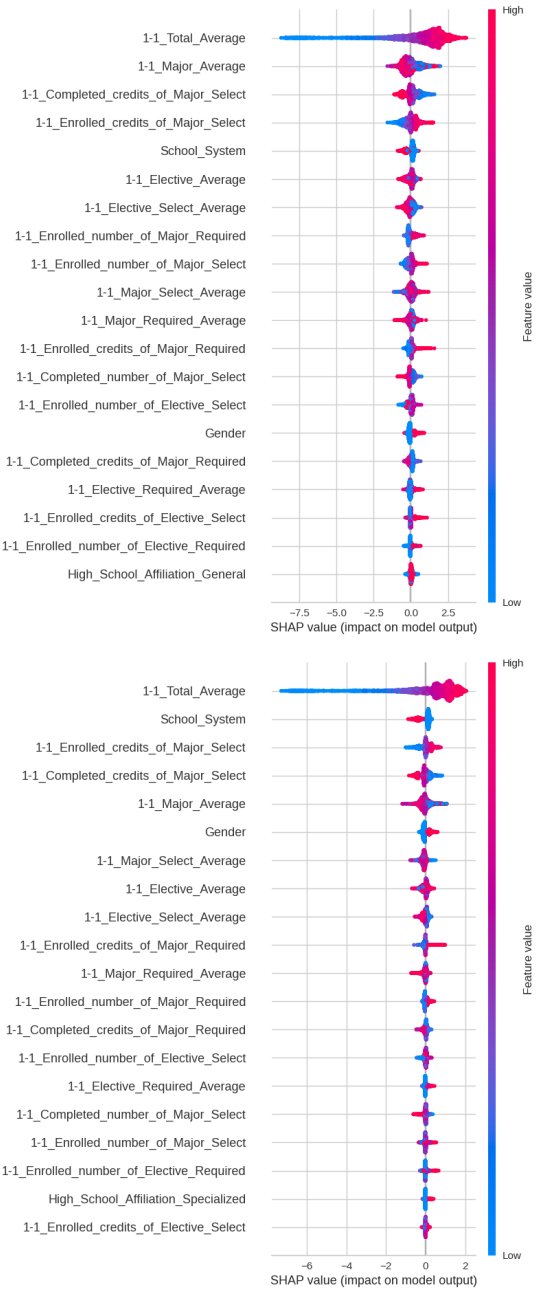

Fig. 3. SHAP value of variables on M1

Fig. 3와 Fig. 4은 XGB(위)와 LGBM(아래)의 제적 확률에 대한 특성들의 영향을 SHAP로 표현한 그림이다. SHAP은 각 알고리즘이 분류모델을 생성하는 데 중요하게 활용된 정도에 따라 개별 변수들을 도표의 위에서부터 아래로 표현하고 있다. 그리고 각 변수들에 대해 표현된 점들의 색상은 타겟 변수를 예측하는 데 어떠한 방향으로 영향을 미쳤는지를 표현하고 있다.

Fig. 3와 Fig. 4에 의하면 M1과 M2에서 공통적으로 제적 및 졸업을 예측하는데 가장 영향을 크게 미친 변수는 1학년 1학기의 총평점평균인데, 해당 점수가 높을수록 졸업을 예측하는데, 그리고 점수가 낮을수록 제적을 예측하는데 주요한 영향을 미쳤다고 해석할 수 있다. 그 중에서도 전공에서의 평점평균이 중요한 영향을 미친 것을 알 수 있다. 더불어, 인구통계적 변수 중 입학 당시의 학제가 M1과 M2에서 일관되게 중요한 변수로 활용되고 있음을 확인할 수 있다.

Fig. 4. SHAP value of variables on M2

위 연구 결과에 따라, 총 전공 평점평균, 총 교양 평점평균, 총 평점평균을 학적별로 상자도표를 그려보면 Fig. 5와 같다. 제적생들의 모든 평점평균 항목에서 낮은 점수를 기록하고 있었다. 또한 휴학생의 성적분포가 재학생의 성적보다 다소 떨어지고 있음을 확인할 수 있었다. 반면, 졸업생과 현재(2023년 기준) 재학생의 성적분포가 매우 흡사함을 알 수 있었다.

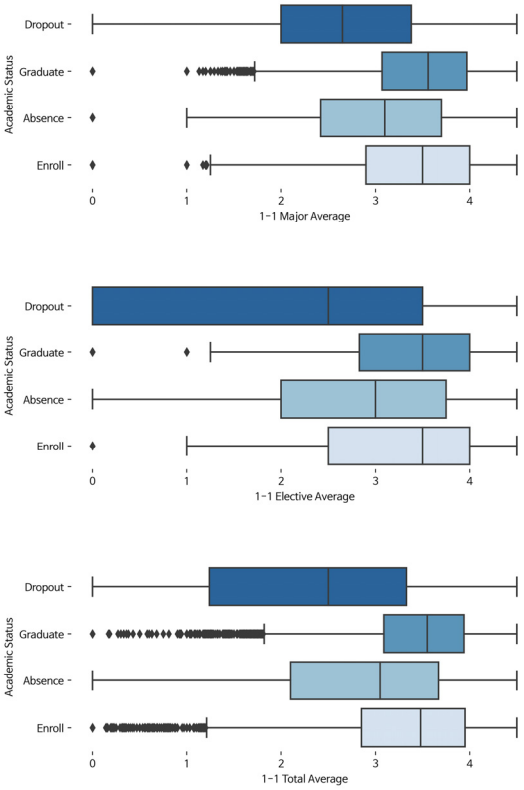

Fig. 5. Academic score by status

5. 결론

본 연구는 서울 소재 전문대학의 학적자료 중에서도 최소한의 인구통계적 변수와 1학년 1학기의 학업성과를 토대로 최종 학적을 예측하는 것을 목표로 하였다. 실험 결과 제적을 예측하는 데 비교적 우수한 성능을 보인 모델은 Boosting 계열의 XGBoost와 LightGBM 모델이었고, 1학년 1학기의 성적정보가 두 모델에서 모두 유의미한 변수로 식별되었다. 이는 첫 학기를 마친 이후의 성적을 토대로 조기에 학생의 미래 학적을 어느 정도 유추할 수 있음을 의미하며, 여러 선행연구에서 강조한 바와 같

이 학업 중단 의도는 학업 시작 초기에 결정된다는 사실을 뒷받침하는 결과이기도 하다[5, 7, 25].

전문대학으로의 진학 결정은 취업을 목표로 하고 있을 가능성이 높다. 가장 최대한 빠른 시기에 취업하고자 하는 동기가 강한 학생들일수록 교육과정의 총기간에 대해 탄력적으로 반응할 수 있을 것이다. 즉, XGBoost와 LightGBM에서의 유의미한 변수를 식별한 결과, 입학당시의 학제(2년제 혹은 3년제)에 따라서도 최종학적에 영향을 미칠 가능성이 있음이 확인하였고, 이는 학업 기간이 학업 중단에 영향을 줄 수 있다는 것으로 풀이된다.

한편, 본 연구에 활용된 데이터는 제적생과 졸업생 간 데이터 불균형 문제가 존재한다. 학적 간 불균형 문제를 고려하여 학업 중단을 예측하고자 한 연구도 일부 존재하나[12, 23], 본 연구에서는 이를 고려하지 않았다. 또한, 초기의 학업 성과를 토대로 예측하는 것이 일부 유의미한 결과를 보이기는 했으나, 분류 성능 수치가 절대적으로 뛰어나다고 하기 어렵다. 이는 학업 성과만으로는 분류모델을 구성하는 데 일부 한계가 있음을 보여준다. 하지만 현재 많은 대학에서 실시하고 있는 대학생활적응검사와 같은 학생들의 심리적, 환경적 요인 등을 파악할 수 있는 검사를 학기 초부터 섬세하게 모니터링하고 이를 활용한다면 더 나은 학업 중단 예측 모델을 구성할 수 있을 것으로 기대한다.

한국의 전문대학들은 조기에 학업 중단 의도를 파악하기 위해 일정 기간 추적과 모니터링이 필요한 정성 자료를 토대로 조기 학업 중단 의도를 파악하는 것은 사실 쉽지 않은 실정이다. 따라서, 본 연구에서는 기존 선행연구에서 주목했던 학생들의 심리적 요인, 사회적 요인, 경제적 요인 등 설문에 의존해야 하는 정성 자료 없이 가장 이른 시기에 파악할 수 있는 학업 성취도 지표를 토대로만 모델링을 시도하였다. 그리고 학업 중단 가능성이 있는 일부 학생들을 필터링하는데 성공하였고, 추후 중퇴에 대한 조기 경보 시스템을 구축하는 데 도움이 될 것이다.

REFERENCES

[1] Agrusti, F., Bonavolontà, G., & Mezzini, M. (2019). University dropout prediction through educational data mining techniques: A systematic review. *Journal of e-learning and knowledge society*, *15(3)*, 161-182. DOI : 10.20368/1971-8829/1135017

[2] Aina, C., Baici, E., Casalone, G., & Pastore, F. (2022). The determinants of university dropout: A review of the socio-economic literature. *Socio-Economic Planning Sciences*, *79*, 101102. DOI : 10.1016/j.seps.2021.101102

[3] Alban, M., & Mauricio, D. (2019). Predicting university dropout through data mining: a systematic literature. *Indian Journal of Science and Technology*, *12(4)*, 1-12. DOI : 10.17485/ijst/2019/v12i4/139729

[4] Ameen, A. O., Alarape, M. A., & Adewole, K. S. (2019). Students' academic performance and dropout predictions: A review. Malaysian *Journal of Computing*, *4(2)*, 278-303. DOI : 10.24191/mjoc.v4i2.6701

[5] Astin, A. W. (1964). Personal and environmental factors associated with college dropouts among high aptitude students. *Journal of Educational Psychology*, *55(4)*, 219. DOI : 10.1037/h0046924

[6] Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. arXiv preprint arXiv:1606.06364. DOI : 10.48550/arXiv.1606.06364

[7] Behr, A., Giese, M., Teguim K, H. D., & Theune, K. (2020). Early prediction of university dropouts-a random forest approach. *Jahrbücher für Nationalökonomie und Statistik*, *240(6)*, 743-789. DOI : 0.1515/jbnst-2019-0006

[8] Bernardo, A. B., Galve-González, C., Núñez, J. C., & Almeida, L. S. (2022). A path model of university dropout predictors: the role of satisfaction, the use of self-regulation learning strategies and students' engagement. *Sustainability*, *14(3)*, 1057.

[9] Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020). Student dropout prediction. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings*, *Part I 21,* 129-140. DOI : 10.3390/su14031057

[10] Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49(4)*, 498-506. DOI : 10.1016/j.dss.2010.06.003

[11] Demeter, E., Dorodchi, M., Al-Hossami, E.,

Benedict, A., Slattery Walker, L., & Smail, J. (2022). Predicting first-time-in-college students' degree completion outcomes. *Higher Education*, 1-21. DOI : 10.1007/s10734-021-00790-9

[12] Lee, S., & Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences*, *9(15)*, 3093. DOI : 10.3390/app9153093

[13] Lizarte Simón, E. J., & Gijón Puerta, J. (2022). Prediction of early dropout in higher education using the SCPQ. *Cogent Psychology*, *9(1)*, 2123588. DOI : 10.1080/23311908.2022.2123588

[14] Lounsbury, J. W., Saudargas, R. A., & Gibson, L. W. (2004). An investigation of personality traits in relation to intention to withdraw from college. *Journal of College Student Development*, *45(5)*, 517-534. DOI : 10.1353/csd.2004.0059

[15] Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, *53(3)*, 950-965. DOI : 10.1016/j.compedu.2009.05.010

[16] Mduma, N., Kalegele, K., & Machuve, D. (2019). A survey of machine learning approaches and techniques for student dropout prediction. *Data Science Journal, 18*, 14-14.

[17] Mortada, L., Bolbol, J., & Kadry, S. (2018). Factors affecting students' performance a case of private colleges in Lebanon. *J Math Stat Anal*, *1*, 105. DOI : 10.5334/dsj-2019-014

[18] Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, *3*, 100066. DOI : 10.1016/j.caeai.2022.100066

[19] OECD. (2019). *Education at a Glance 2019 : OECD Indicators.* Paris : OECD Publishing. DOI : 10.1787/19991487

[20] Orooji, M., & Chen, J. (2019, December). Predicting louisiana public high school dropout through imbalanced learning techniques. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)* (pp. 456-461). IEEE. DOI : 10.1109/ICMLA.2019.00085

[21] Shafiq, D. A., Marjani, M., Habeeb, R. A. A., & Asirvatham, D. (2022). Student retention using educational data mining and predictive analytics: a systematic literature review. *IEEE Access*. DOI : 10.1109/ACCESS.2022.3188767.

[22] Tan, M., & Shao, P. (2015). Prediction of student dropout in e-Learning program through the use of machine learning method. *International journal of emerging technologies in learning*, *10(1)*. DOI : 10.3991/ijet.v10i1.4189

[23] Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, *41(2)*, 321-330. DOI : 10.1016/j.eswa.2013.07.046

[24] Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, *45(1)*, 89-125. DOI : 10.2307/1170024

[25] Tinto, V. (2006). Research and practice of student retention: What next?. Journal of college student retention: *Research, Theory & Practice*, *8(1)*, 1-19. DOI : 10.2190/4YNU-4TMB-22DJ-AN4W

[26] Young A, Song., Sinae, Kim. (2019). Factors Affecting College Freshmen's Intention to Drop Out. *The Korea Contents Association*, *19(6)*, 257-270. DOI : 10.5392/JKCA.2019.19.06.257

[27] Youngsik, Woo., Minok, Song. (2022). Relationship Between Carrer Decision Level, Academic Self-effcacy, Self-directed Learning Ability, and College Life Adptation of Junior College Freshmen. *The Journal of Humanities and Social Science 21*, *13(4)*, 1417-1432.

[28] Kostat. (2023). *Birth Statistics in 2022* [Press Release].

오 상 조(Sangjo Oh) [정회원]

- 1991년 2월 : 서울대학교 경영학과 (경영학 학사)
- 1995년 2월 : 서울대학교 경영학과 (경영학 석사)
- 2002년 2월 : 서울대학교 경영학과 (경영학 박사)
- 1996년 3월~현재 : 동양미래대학교 경영정보학과 교수
- 관심분야 : 경영정보, ebusiness, 데이터사이언스
- E-Mail : san@dongyang.ac.kr

심 지 환(JiHwan Sim) [정회원]

- 2017년 2월 : 국민대학교 회계학과 경영학석사
- 2021년 9월 : 국민대학교 데이터사인스학과 경영학 박사수료
- 2022년 9월~현재 : 동양미래대학교 경영정보학과 조교수
- 관심분야 : 기계학습, 데이터사이언스
- E-Mail : sim2080@dongyang.ac.kr