대용량 LMS 로그 데이터를 이용한 심층신경망 기반 대학생 학업성취 조기예측 모델

The Early Prediction Model of Student Performance Based on Deep Neural Network Using Massive LMS Log Data

문기범, 김진원, 이진숙
고려대학교 디지털정보처

Kibum Moon(a072826@korea.ac.kr), Jinwon Kim(herojin333@korea.ac.kr),
Jinsook Lee(gupye@korea.ac.kr)

요약

학습관리 시스템(LMS)에 축적되는 로그 데이터는 학습 과정에 대한 양질의 정보를 제공한다. 지금까지 LMS 로그 데이터를 활용한 학업성취 예측 연구가 다양하게 수행되었지만, 상대적으로 적은 양의 학생 및 수업 데이터에 기반하고 있어 연구 결과 일반화 가능성에 한계가 존재한다. 본 연구는 대용량 LMS 로그 데이터를 이용해 대학생 학업성취를 조기예측하는 심층신경망 모델을 개발하고 성능을 검증했다. 이를 위해 가명화 처리된 LMS 로그 데이터 78,466,385건과 성적 데이터 165,846건을 활용했다. 그 결과, 본 연구에서 제안하는 예측모델은 우수학생 집단을 학기 초부터 높은 수준의 정확도로 예측하였다. 한편 보통 및 저성취 집단에 대한 예측 정확도는 제한적인 수준이었지만, 예측시점이 늦을수록 향상되었다. 본 연구의 결과는 순수 LMS 로그 데이터만을 이용해 실제로 활용할 수 있을 정도의 일반화 성능을 가진 심층신경망 기반 조기예측 모델을 구현했다는 의의가 있다.

■ 중심어 : | 학습관리시스템 | 학생성공 | 인공지능 | 빅데이터 | DNN

Abstract

Log data accumulated in the Learning Management System (LMS) provide high-quality information for the learning process of students. Until now, various studies have been conducted to predict students' academic achievement using LMS log data. However, previous studies were based on relatively small sample sizes of students and courses, limiting the possibility of generalization. This study developed and validated a deep neural network model for the early prediction of academic achievement of college students using massive LMS log data. To this end, we used 78,466,385 cases of LMS log data and 165,846 cases of grade data. The proposed model predicted the excellent-grade students with a high level of accuracy from the beginning of the semester. Meanwhile, the prediction accuracy for the moderate and underachieving groups was relatively low, but the accuracy improved as the time points of the prediction were delayed. This study is meaningful in that we developed an early prediction model based on a deep neural network with sufficient accuracy for practical utilization by only using LMS log data.

■ keyword : | Learning management system | Student success | Artificial intelligence | Big data | DNN

I. 서 론

학습관리 시스템(Learning Management System, 이하 LMS)은 학습자의 학습을 관리하고 지원하는 온라인 시스템을 말한다. COVID-19의 대유행으로 대면 수업이 제한되고 대부분의 강의가 온라인으로 진행됨에 따라[1], 대학 교육 내 LMS의 역할과 중요성이 과거보다 크게 높아졌다[2]. LMS는 학습 콘텐츠 확인, 강의영상 재생, 글쓰기, 퀴즈 응시 등 학습자의 모든 행동에 대한 로그 데이터를 저장하기 때문에, 학습자의 학습 과정을 면밀하게 검토할 수 있는 양질의 정보를 제공한다[3][4]. 최근 인공지능 기반 데이터

분석 기법의 발전과 개인화된 교육의 중요성이 대두되면서 LMS를 활용한 학습자 행동 및 학습참여 분석[5-8], 학업성취 예측[4][9-15], 학습동기 강화 및 개입[16] 등 영역에 대한 연구가 활발하게 진행되고 있다.

특히, LMS 로그 데이터를 활용한 학업성취 조기예측은 학생들의 중도탈락을 예방하고 학생성공 가능성을 높이는 측면에서 중요한 함의를 갖는다[4][14][15]. 예를 들어, 학기 초에 학업적 어려움을 겪는 학생을 조기에 식별할 수 있다면 학업적 성과를 개선하는 데 더 많은 시간을 사용할 수 있다. 마찬가지로, 우수한 성적을 받을 것으로 예측되는 학생들에게 해당 정보를 제공함으로써 추가적인 동기를 부여할 수 있다.

지금까지 LMS 로그 데이터를 활용한 연구는 주로 통계적 방법론을 사용해 LMS 이용과 학업성취 사이의 관계를 탐색했다[7][9][10][12]. 그런데, 통계적 방법론을 사용해 도출한 모델은 기존 데이터를 잘 설명할 수 있는 반면, 새로운 데이터에 대한 예측력이 보장되지 않는다는 한계를 가지고 있다[17]. 최근에는 LMS 로그 데이터와 기계학습(machine learning) 방법론을 활용해 학생의 학업성취를 예측하고자 하는 시도가 다양하게 전개되고 있다[4][11][13].

기계학습 기반 학업성취 예측모델을 실제 대학의 교육 서비스에 활용하기 위해서는 예측모델의 성능뿐만 아니라 일반화(generalization) 가능성이 보장되어야 한다. 만약 연구에서 제안한 예측모델이 소수의 과목이나 학생에 대해서만 우수한 성능을 보인다면 실제 서비스로써 활용 가치가 제한적일 것이다. 하지만, 지금까지 수행된 LMS 로그 데이터 기반 학업성취 예측연구는 수백에서 천명 정도 수준의 적은 수의 학생 데이터에 기반해 모델 제작 및 검증이 이루어졌기 때문에 예측모델의 일반화 가능성을 확인하는 데 한계가 존재한다. 특히 지금까지 보고된 연구 중 학과 및 수업과 무관하게 대학 전체의 데이터를 활용한 연구는 보고된 적 없는 것으로 보인다.

따라서, 본 연구에서는 전체 대학 수준의 대용량 LMS 로그 데이터와 성적 데이터를 활용해 기계학습 기반 예측모델을 개발하고 성능을 검증하고자 한다. 이를 통해 제안하는 모델이 실제 교육 장면에서 활용할 수 있을 정도의 일반화 성능을 가졌는지 확인하고자 했다. 예측모델을 위한 알고리즘으로 데이터의 양이 많을수록 높은 성능을 보이는 것으로 알려진 심층신경망 알고리즘을 활용했다. 본 연구에서 제안하는 학업성취 예측모델이 조기 예측력을 가졌는지 탐색하기 위해 학기 전체 데이터뿐만 아니라 학기가 각각 25%, 50%, 75% 경과한 시점의 데이터만을 활용해 학업성취를 예측하는 모델을 추가로 제작했다. 마지막으로, LMS 로그 데이터의 고유한 예측력을 확인하기 위해 LMS 로그 데이터 외에 중간 및 기말고사 성적, 퀴즈 결과, 과제물 평가 점수 등 학업성취와 관련된 변인을 모델에 반영하지 않았다.

II. 방 법

1. 연구 대상 및 절차

본 연구에서는 학기 경과 시점별 LMS 로그 데이터를 활용해 학업성취를 예측하는 심층신경망 모델을 개발하고 성능을 검증하고자 한다. 본 연구를 위해 서울 소재 한 사립대학 본교(20,142명, 76.50%) 및 분교(6,189명, 23.50%) 소속 학생 26,331명의 2020년 1학기 LMS 로그 데이터 78,466,385건과 과목별 성적 데이터 165,846건을 추출했다. LMS 로그 데이터 외에도 학생별 입학성적, 소속 캠퍼스, 학과, 과목별 개설학과 등의 정보는 학업성취를 예측하는데 중요한 정보를 제공할 수 있다. 하지만, 모델에 투입되는 변수가 많아질수록 모델의 확장성과 일반화 가능성이 작아진다. 본 연구에서는 LMS 로그 데이터만을 사용함으로써 확장성(scalability)과 일반화 가능성(generalizability)이 높은 모델을 개발하고자 했다. 이러한 목적에 따라 LMS 로그 데이터 외에 어떤 추가적인 학생 정보나 과목 정보도 모델 개발 및 훈련 과정에 활용하지 않았다. 또 시험 및 퀴즈 성적, 과제물 평가 점수 등도 LMS 서버에 저장되지만, 그 자체로 학업성취와 밀접하게 관련된 변인들은 분석에서 제외했다. 이러한 제한을 통해 학업성취에 대한 LMS 로그의 독립적인 설명력을 확인하고, 다양한 대학에 적용할 수 있는 확장성과 일반화 가능성을 갖춘 모델을 개발하고자 하였다.

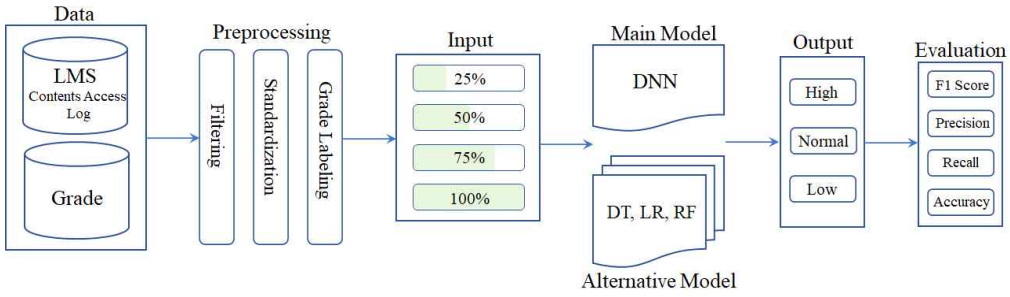

그림 1. 연구절차

개인정보 보호를 위해 LMS 로그 데이터와 성적 정보에 대한 가명화 처리를 진행했다. LMS 로그 데이터와 과목별 성적 데이터를 병합하기 위해 개인을 식별할 수 없도록 학생마다 알파벳과 숫자로 이루어진 임의의 식별자를 부여했다. 가명화 작업 중 데이터가 유출되지 않도록 외부망과 독립된 내부망에서 모든 가명화 작업을 진행했다. 두 데이터 세트의 병합이 끝난 후 가명 식별자를 삭제했다. 이후 분석 절차에서는 학생 개인정보와 식별자가 사용되지 않기 때문에 학생 식별에 따른 잠재적인 위험이 크지 않을 것으로 예상된다.

2. 데이터 추출 및 전처리

2.1 과목

예측 모델의 안정성을 위해 수강인원 수가 30명 이상인 강의만을 분석에 활용했다. 이때, 같은 과목이더라도 분반이 다르면 다른 과목으로 간주했다. 성적을 예측하는 것이 모델의 목표이므로 성적이 주어지지 않는 P/F (Pass or Fail) 과목은 분석에서 제외했다. 이수학점 수에 따라 요구되는 주당 학습 시간이 상이하기 때문에 3학점 강의만 모델 학습 및 검증에 사용했다. 상기 절차를 거친 3,888건의 수업을 분석에 활용했다.

2.2 LMS 로그 데이터

LMS 로그 테이블에는 학생별 LMS 시스템 이용 내역이 이벤트 타입에 따라 시간과 함께 기록된다. 이벤트의 종류로는 로그인, 로그아웃, 과목별 콘텐츠 접근, 세션 시작, 세션 끝 등이 있다. 본 연구에서는 상기 이벤트 종류 중에서 수업 별로 집계되는 과목별 콘텐츠 접근 로그 데이터를 활용했다. 콘텐츠 접근 로그는 학생이 해당 과목의 공지 및 수업자료 확인, 온라인 콘텐츠 시청, 퀴즈 응시, 글 작성 등의 학습활동을 할 때마다 저장된다. 본 연구에서는 LMS 로그 데이터에 기반 학업성취 예측 시 사용 시간보다 접속 횟수가 중요하다는 선행연구 결과를 고려해 접속 횟수 데이터만을 산출했다[18]. 또, 모델의 간결성을 위해 학생이 접근한 콘텐츠의 유형은 고려하지 않았다.

추출한 LMS 로그 데이터에 대한 전처리는 다음과 같이 진행했다. 먼저, 학생-강의 쌍별 일일 콘텐츠 접근 횟수를 집계했다. 이상치를 제거하기 위해 과목별 일일 콘텐츠 접근 횟수가 2,000회를 초과한 기록이 있는 9명의 LMS 데이터를 분석에서 제외했다. 이상치 처리 후 일일 콘텐츠 접근 횟수의 평균과 표준편차는 각각 10.00, 15.41이었다.

LMS 활용 패턴은 과목과 날짜에 따라 다를 수 있다. 예를 들어, 교수자가 제공하는 콘텐츠가 많은 수업은 그렇지 않은 수업에 비해 수강생의 평균 콘텐츠 접속 횟수가 더 높을 것이다. 비슷하게, 주말보다 평일에 콘텐츠 접속 횟수가 많을 것이다. 이러한 차이를 통제하기 위해 일일 콘텐츠 접근 횟수에 대한 표준화를 실시했다. 표준화를 거친 일일 콘텐츠 접근 횟수는 해당 콘텐츠 접근 횟수가 같은 날 같은 과목을 수강하는 다른 학생에 비해 얼마나 많거나 적은지를 나타낸다. 상기 전처리 과정을 거친 일일 콘텐츠 접근 횟수의 표준화 값을 학업성취 예측모델의 입력 데이터로 사용했다.

2.3 학업성취

과목별 성적 등급을 학업 성취 지표로 사용했다. 연구 대상 대학의 성적 등급은 A+, A, B+, B, C+, C, D+, D, F의 8단계로 구성되어 있다. 본 연구에서는 8단계를 다시 크게 세 집단으로 나누었다. 구체적으로, 4.5와 4는 우수, 3.5부터 2는 보통, 1.5, 1, 0은 저성취 집단으로 구분했다. 전체 과목별 성적 데이터 중 우수, 보통, 저성취 집단이 차지하는 비율은 각각 56.27%, 38.68%, 5.05%였다.

3. 예측모델 개발

본 연구에서는 상기 전처리를 거친 LMS 로그 데이터로 세 가지 학업성취 집단을 분류하는 예측모델을 개발하고자 한다. 분류 알고리즘으로 전연결 심층신경망(Fully Connected Deep Neural Network, 이하 DNN)을 활용했다. DNN은 입력층(input layer), 출력층(output layer), 그리고 두 개의 층 사이에 여러 개의 은닉층(hidden layer)이 존재하는 인공신경망(Artifitial Neural Network, ANN)을 말한다. 모델의 조기예측 성능을 검증하기 위해 예측 시점을 네 단계로 구분했다. 즉, 학기 경과 28일(25%), 55일(50%), 82일(75%), 103일(100%) 시점까지의 LMS 로그 데이터로 학업성취를 예측하는 모델을 각각 개발했다. 예측 시점별 DNN 모델의 분류 성능을 일반적인 분류 알고리즘인 로지스틱 회귀모델, 의사결정 나무, 랜덤 포레스트 등과 비교했다.

3.1 모델 구조

그림 2는 본 연구에서 적용한 DNN 알고리즘의 구조를 나타낸다. 입력층에 포함된 노드의 수는 예측 시점별 학기 경과일수에 대응하는 28, 55, 82, 103개다. 은닉층의 수를 결정하기 위해 은닉층의 수를 0개에서 5개까지 늘려가며 성능을 비교했다. 그 결과, 은닉층의 수가 2개일 때 최적의 학습효율과 분류 성능을 보였다. 따라서 2개의 은닉층을 가진 모델을 선택했다. 각 노드의 활성화 함수로 ReLU (Rectified Linear Unit)를 사용했다. 출력층에서는 은닉층으로부터 정보를 받아 우수, 보통, 저성취 집단 각각에 대한 예측치를 산출한다. 따라서 출력층의 노드가 3개가 되도록 설정했다. 결과예측을 위한 활성화 함수로 Softmax를 사용했다. 예측 결과와 실제 값 사이의 오차를 계산하는 손실함수로 cross entropy를 사용했으며, 최적화 기법은 Adam을 사용했다. 과적합 문제를 막기 위해 레이어 마다 배치정규화 단계를 추가하고, 마지막 은닉층 뒤에 dropout 층을 추가해 학업성취 집단 예측 시 신경망 일부가 사용되지 않도록 만들었다.

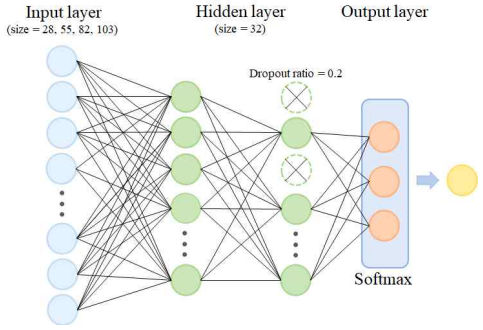

그림 2. DNN 알고리즘

3.2 데이터 분할

홀드아웃 방법(holdout method)을 사용해 전처리가 끝난 데이터 세트의 90%는 훈련 데이터 세트로, 나머지 10%는 테스트 데이터 세트로 구분했다. 모델의 과적합을 방지하고 일반화 가능성을 높이기 위해 테스트 데이터는 최종 선택된 모델의 성능을 검증하는 목적으로만 사용했으며, 모델 훈련, 초매개변수 최적화, 모델 선택 등의 절차는 훈련 데이터 세트를 활용했다.

표 1은 두 데이터 세트 내 학업성취 집단별 빈도를 보여준다. 학업성취 집단 간 불균형이 존재하기 때문에 층화추출(stratified sampling)을 사용해 두 데이터 세트에서 각 학업성취 집단이 차지하는 빈도가 일정하도록 만들었다. 데이터 불균형을 처리하기 위해 훈련 데이터 세트에 대해 다수 집단의 빈도에 맞춰 소수 집단 표본을 생성하는 오버 샘플링(oversampling) 기법을 사용했다. 오버 샘플링 기법으로 무작위로 소수 집단 표본을 생성하는 ROS(Random Over Sampler), k-최근접(k-Nearest Neighbors) 이웃 알고리즘에 기반해 가장 근접한 소수 집단 표본 데이터 사이의 직선 위 임의의 위치에 가상 표본을 생성하는 SMOTE(Synthetic Minority Over-sampling) 기법, 그리고 기존 SMOTE를 개선해 경계선에 존재하는 데이터만을 사용해 SMOTE를 수행

하는 Borderline SMOTE 기법을 사용했다. 세 방법 모두 데이터 불균형으로 인한 낮은 재현율(recall) 문제를 어느 정도 개선했지만, 정밀도(precision)와 정확도(accuracy)에서의 손실이 이보다 큰 것으로 나타났다. 따라서 이후 분석에서는 세 가지 방법 모두를 반영하지 않았다.

표 1. 데이터 세트 내 학업성취 집단별 빈도(백분위)

데이터 세트	우수	보통	저성취	합계
훈련 데이터	54,474 (56.27%)	37,451 (38.68%)	4,891 (5.05%)	96,816 (100%)
테스트 데이터	6,053 (56.27%)	4,162 (38.69%)	543 (5.05%)	10,758 (100%)

3.3 초매개변수 탐색 및 모델 선택

최적의 모델을 식별하기 위해 초매개변수 탐색 절차를 수행했다. 이 절차에서는 모델 초매개변수 조합을 다양하게 변형해가며 가장 좋은 성능을 보이는 모델을 찾는다. 성능 측정의 신뢰도를 높이기 위해 층화 k겹 교차검증(Stratified k-fold cross validation) 기법을 사용했다. k겹 교차검증은 가용한 데이터를 훈련과 검증에 활용할 수 있으며, 과적합이 발생할 가능성을 낮춰준다는 장점이 있다. 하지만 k값이 커질수록 모델 훈련에 긴 시간이 소요된다. 본 연구에서는 k값을 3으로 설정했다.

본 연구에서는 Epoch 횟수, 배치 크기, 학습률, 각 층별 노드의 수, dropout 비율을 조정했다(표 2). 네 가지 예측시점 모델별로 각각 108개 조합에 대한 grid 탐색을 실시한 후, cross entropy loss가 가장 작은 모델을 선택했다. 표 3은 각 예측시점별 최종 모델에 사용된 파라미터 조합을 보여준다. 상기 절차를 통해 최종 선택된 각 모델별 초매개 변수를 표 2에 제시했다.

표 2. 초매개변수 탐색에 사용된 매개변수

매개변수	값
Epoch 횟수	[50, 100, 200]
배치 크기	[4096, 8192]
은닉층 크기	[32, 64, 128]
학습률	[0.001, 0.0001]
Drop 비율	[0, 0.1, 0.2]

표 3. 예측시점별 최종모델 초매개변수

	초매개변수					
예측 시점	Input 크기	Epoch 횟수	배치 크기	은닉층 크기	학습률	Dropout 비율
25%	28	50	4096	32	0.001	0.2
50%	55	50	4096	32	0.001	0.2
75%	82	200	4096	32	0.0001	0.2
100%	103	200	4096	32	0.0001	0.2

3.4 대안모델

본 연구에서 제안하는 DNN 기반 예측 모델 외에 분류 과제에서 널리 활용되는 Lasso Logistic Regression(LR), Decision Tree(DT), Random Forest(RF)를 대안 모델로 적용했다. 상기 대안모델 또한 DNN 모델과 마찬가지로 층화 k겹 교차검증을 활용한 초매개변수 탐색 및 모델선택 과정을 거쳤다.

3.5 성능 평가

예측모델이 선택되면 테스트 데이터 세트를 활용해 최종 모델의 분류성능을 평가했다. 모델의 일반화 가능성을 높이기 위해 본 절차에 사용될 테스트 데이터 세트는 최종 모델성능 평가 목적으로만 사용했다. 성능평가 지표로는 분류성능을 평가하는 데 널리 활용되는 Precision, Recall, F1-Score, Accuracy를 사용했다. 각 지표의 계산식은 다음과 같다.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Precision, Recall, F1-Score은 우수, 보통, 저성취

집단 분류 각각에 대해 평가했으며, Accuracy는 모델의 전반적인 성능을 평가하기 위해 집단을 가리지 않고 전체 테스트 세트에 대해 한번 평가했다.

3.7 실험 환경

예측모델 제작 및 검증을 위해 Python 3.8[19]의 Pytorch 1.8.0[20], Skorch 0.10.0[21], Scikit-learn 0.24.1[22] 라이브러리를 사용했다. 본 연구에 사용된 컴퓨팅 자원은 다음과 같다. Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz(20 cores), GeForce RTX 2080 SUPER, 394.8594 GB DDR RAM, Ubuntu 20.04.2 LTS OS.

III. 결 과

예측시점별 LMS 로그 데이터 기반 대학생 학업성취 예측모델의 성능을 평가하기 위한 실험을 진행했다. 이를 위해 앞선 모델 훈련 및 선택 절차에서 한 번도 사용되지 않은 테스트 데이터 세트를 사용해 모델의 성능을 시험했다. 표 4의 혼동 행렬(confusion matrix)은 모델이 예측한 결과와 실제 결과를 보여준다. 표 5는 각 시점별 모델의 성능지표를 나타낸다. 전반적으로 예측 시점이 학기 말에 가까울수록 모델의 성능이 좋은 것으로 나타났다. 하지만. 학기 경과 100% 시점의 Accuracy와 25% 시점의 Accuracy의 차이는 0.017로 크지 않았다.

LMS 로그 데이터를 활용한 학업성취 예측 모델이 가장 정확하게 분류해낸 집단은 우수집단이었다. 우수집단에 대한 예측은 모든 시점에서 F1-Score 기준 0.7 이상의 정확도를 보였다. 특히, 학기 경과 25% 시점의 데이터만을 사용했을 때도 우수집단 분류의 Recall 값은 0.92 수준으로 매우 양호하게 나타났다. 이는 예측모델이 실제로 A+ 또는 A학점을 받은 학생의 92%를 우수집단으로 분류했음을 의미한다.

반면, F1-Score를 기준으로 보았을 때, 우수집단 예측보다 보통 및 저성취 집단에 대한 예측의 성능은 상대적으로 낮았다. 보통 집단에 대한 예측의 Precision은 전 예측시점에 걸쳐 0.47 ~ 0.50이었으며, Recall은 0.14 ~0.24였다. 저성취 집단에 대한 예측의 Recall은 0.09 ~0.24로 세 집단 중 가장 저조했던 것에 반해, Precision은 0.62 ~ 0.69로 오히려 우수집단에 대한 예측보다 우수한 것으로 나타났다. 이는 예측모델이 저성취 집단으로 예측한 학생의 약 62%에서 69%가 실제로 D 또는 F학점을 받았다는 것을 의미한다.

표 4. 테스트 데이터 세트의 혼동 행렬

예측시점	실제값	예측값		
		우수	보통	저성취
25%	우수	5,546	501	6
	보통	3,569	577	16
	저성취	346	148	49
50%	우수	5,483	562	8
	보통	3,446	687	29
	저성취	282	191	70
75%	우수	5,311	732	10
	보통	3,245	869	48
	저성취	257	191	95
100%	우수	5,245	797	11
	보통	3,133	979	50
	저성취	214	196	133

다음으로, 본 연구에서 제안하는 DNN 기반 모델과 LR, DT, RF를 사용한 대안모델의 성능을 비교했다 (그림 3). 그 결과. 우수 집단에 대한 예측에서는 DNN 모델과 대안모델 사이의 차이가 크지 않은 것으로 나타났다. 세 가지 대안 모델 모두 전 예측시점에 걸쳐 F1-Score 0.71 이상의 양호한 성능을 보였다. 하지만 보통집단과 저성취 집단에 대한 예측에서는 DNN모델이 나머지 대안모델에 비해 모든 예측시점에서 더 우수한 예측성능을 나타냈다. 특히 저성취 집단에 대한 예측에서 DNN모델은 학기가 경과할수록 더 정확한 예측을 하는 데 반해, 세 가지 대안모델 모두 이러한 성능향상을 보이지 않았다.

표 5. DNN기반 모델 성능지표

예측시점	학업성취 집단	Precision	Recall	F1-score	Accuracy
25%	우수	0.59	0.92	0.71	0.57
	보통	0.47	0.14	0.21	
	저성취	0.69	0.09	0.16	
50%	우수	0.60	0.91	0.72	0.58
	보통	0.48	0.17	0.25	
	저성취	0.65	0.13	0.22	
75%	우수	0.60	0.88	0.71	0.58
	보통	0.48	0.21	0.29	
	저성취	0.62	0.17	0.27	
100%	우수	0.61	0.87	0.72	0.59
	보통	0.50	0.24	0.32	
	저성취	0.69	0.24	0.36	

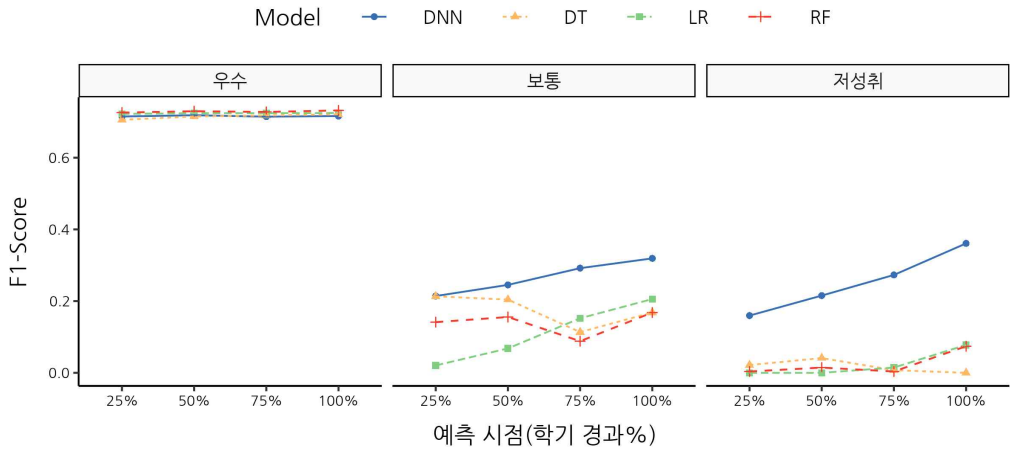

그림 3. 모델간 성능 비교. DNN = 심층신경망(Deep Neural Network); DT = 의사결정나무(Decision Tree); LR = Lasso 로지스틱 회귀(Lasso Logistic Regression); RF = 랜덤 포레스트(Random Forest)

IV. 논 의

본 연구에서는 대용량 LMS 로그 데이터를 활용해 심층신경망 기반 대학생 학업성취 예측모델을 제작하고 성능을 검증했다. 전체 학기 데이터를 이용하는 대신 학기 경과 시점별 데이터를 이용함으로써 학업성취에 대한 조기예측이 가능한지 탐색했다. 또, LMS 과목 콘텐츠 조회 빈도 데이터만을 사용함으로써 학업성취에 대한 LMS 로그 데이터의 독립적인 예측력을 확인했다. 그 결과, 본 연구에서 제안하는 심층신경망 기반 학업성취 조기예측 모델은 양호한 수준의 예측성능과 일반화 가능성을 보이는 것으로 나타났다. 구체적으로, 예측모델은 우수학생 집단을 가장 잘 예측하였으며, 학기 초부터 높은 수준의 정확도를 보였다. 한편, 보통 및 저성취 집단에 대해서는 상대적으로 제한된 수준의 정확도를 보였지만, 예측시점이 늦어질수록 예측의 정확도가 향상되었다.

본 연구의 결과는 선행연구와 일치되게 LMS 로그 데이터를 이용한 예측모델이 대학생 학업성취를 확률보다 높은 수준으로 예측할 수 있다는 것을 보여준다[4][11][13]. 본 연구는 대량의 LMS 로그 데이터 및 성적 데이터를 이용함으로써 기계학습 기반 학업성취 예측모델의 일반화 가능성을 한 대학 수준에서 검증했다는 의의가 있다. 즉, 가용한 데이터의 제한에 따라 서로 다른 수업에 대한 예측 정확도가 제한적이었던 선행연구와 달리[14], 본 연구의 결과에 제시된 모델의 성능 지표는 특정 과목이나 학생에 대해서 뿐만 아니라, 전체 대학에 적용했을 때 예상되는 성능을 나타낸다.

본 연구의 또 다른 의의는 LMS 로그 데이터만을 이용해 학생의 학업성취를 조기에 예측할 수 있는 가능성을 확인했다는 데 있다. 본 연구의 결과는 본 연구에서 제안하는 심층신경망 모델뿐만 아니라 세 가지 대안모델 모두 학기 초 데이터만으로도 우수학생을 비교적 정확하게 예측할 수 있다는 것을 보여준다. 교육 서비스 영역에서 학업성취 예측모델이 필요한 이유는 단지 학생을 학업성취에 따라 정확하게 분류하기 위해서가 아니라 학생이 더 나은 성과를 낼 수 있게 동기를 부여하거나 필요한 개입을 조기에 제공할 수 있게 만들기 때문이다[4][14][15]. 만약 학생의 성적을 100%의 정확도로 예측하는 모델이 있더라도 전체 학기의 데이터를 모두 사용해야 한다면, 이 모델은 학생성공 지원이라는 교육적 목적의 측면에서는 활용도가 낮을 것이다. 이런 점에서 본 연구의 결

과는 추후 개인화된 교육 서비스를 개발하고 적용하는 데 LMS 로그 데이터와 이를 활용한 기계학습 예측모델이 중요한 역할을 수행할 수 있다는 함의를 제공한다.

또, 본 연구는 LMS 로그 데이터 외에 중간 및 기말고사 성적, 퀴즈 결과, 과제물 평가 점수 등 학업 성취 관련 변인을 활용한 기존 연구와 다르게 [4][11][13][14], 순수한 LMS 로그 데이터만을 사용함으로써 LMS 로그 데이터의 고유한 예측력을 확인했다. 특히 본 연구에서 수행한 LMS 로그 데이터에 대한 전처리 절차가 예측모델의 예측력을 높이는 데 기여했음을 암시한다. 즉, 한 학생이 특정 과목을 수강하면서 해당 과목의 콘텐츠에 얼마나 자주 접근하는지를 같은 날 같은 과목을 수강하는 다른 학생과 비교한 값은 단순한 출석 정보나 콘텐츠 접근 빈도보다 학업성취를 예측하는 데 더 의미있는 정보를 제공하는 것으로 이해할 수 있다.

본 연구에서 제안하는 연구결과 활용방안은 아래와 같다. 먼저, 가장 정확한 성능을 보인 우수집단 예측결과를 LMS나 이메일 등을 통해 제공함으로써 학생의 학습동기를 높일 수 있을 것이다. 한편, 저성취 집단에 대한 예측의 경우, 전반적으로 낮은 F1-Score에도 불구하고 저성취 학생에 대한 학업지원 측면에서 중요한 함의를 갖는다. 본 연구에서 제안하는 모델은 F 또는 D 학점을 받는 학생의 10% ~ 20%만을 정확하게 예측하지만, 저성취 집단으로 예측되는 학생 중 65% ~ 69%는 실제로 D나 F 학점을 받는 것으로 나타났다. 이러한 수치는 교수자 혹은 학생지원 주무 부처에서 저성취 집단으로 예측되는 학생에 대한 선제적 개입을 고려해 볼 수 있는 근거가 될 수 있다.

본 연구의 첫 번째 한계는 다음과 같다. 본 연구에서 제안하는 예측모델은 우수학생에 대해서는 전 시점에 걸쳐 비교적 정확하게 식별하는 반면, 저성취 학생 식별에서는 제한된 수준의 성능을 보였다. 이러한 한계에는 세 가지 원인이 있을 수 있다. 먼저, 성적 데이터 자체의 불균형으로 인해 저성취 집단에 대한 충분한 학습이 이루어지지 않거나, 예측모델의 편향이 발생했을 가능성이 있다. 또 다른 가능성으로 우수학생의 경우 모든 과목에 걸쳐 전반적으로 비슷한 LMS 로그 패턴을 보이는 반면, 저성취 집단의 경우 학생마다 그리고 과목마다 더 다양하고 이질적인 패턴을 보일 가능성이 있다. 마지막으로, 본 연구에서 사용한 콘텐츠 접근 로그 데이터의 경우 단순 빈도만 사용하였기 때문에 같은 행위에 대한 질적인 차이를 반영하지 못한다. 예컨대, 한 콘텐츠를 확인한 후 다음 콘텐츠를 확인하는데 소요되는 시간이 저성취 집단의 학생은 우수 및 보통 집단 학생보다 평균적으로 짧을 가능성이 있다. 즉, 시간적인 측면을 추가로 고려함으로써 개별 콘텐츠를 얼마나 자세하게 탐색하는지 습관적인 부분을 추가로 고려할 수 있을 것이다. 비슷한 맥락에서 각 콘텐츠의 성격에 따라 콘텐츠를 조회하는 순서도 학생들의 학습활동에 대한 추가적인 정보를 제공할 수 있을 것이다. 따라서 후속 연구에서는 저성취 집단에 대한 학습이 충분하게 이루어질 수 있도록 더 많은 양의 균형 잡힌 데이터를 활용해야 할 것이다. 또, 로그 데이터의 시간이나 순서 같은 요소를 추가로 고려함으로써 조기예측 모델의 저성취 집단에 대한 예측력을 높일 수 있을 것이다.

본 연구의 두 번째 한계는 다음과 같다. 본 연구가 기존 연구들보다 더 많은 양의 LMS 로그 데이터와 성적 데이터를 활용했지만, 하나의 대학에서 수집한 한 학기 자료라는 한계가 있다. 이로 인해, 본 연구의 결과가 다른 대학에서 수집된 데이터나, 다른 시점에 수집된 데이터에 대해서도 나타날 수 있는지 아직 검증되지 않았다. 따라서 LMS 로그 데이터를 이용한 기계학습 기반 예측모델의 일반화 가능성과 확장성을 더욱 폭넓게 탐색하기 위해서는 다양한 대학에서 여러 학기에 걸쳐 수집한 데이터를 활용한 연구가 수행되어야 한다.

참 고 문 헌

[1] 교육부, “모든 학생을 위한 원격교육 환경 구축에 총력 - 원격교육 환경 구축을 위해 교육부·과학기술정보통신부 힘 모으기로 -,” 보도자료, 2020.04.01.

[2] 이용상, 신동광, “코로나 19로 인한 언택트 시대의 온라인 교육 실태 연구,” 교육과정평가연구, 제23권, 제4호, pp.39-57, 2020.

[3] E. W. Black, D. Beck, K Dawson, S. Jinks, and M. DiPietro, "The other side of the LMS: Considering

implementation and use in the adoption of an LMS in online and blended learning environments," TechTrends, Vol.51, No.2, pp.35-53, 2007.

[4] M. Riestra-González, M. del Puerto Paule-Ruíz, and F. Ortin, "Massive LMS log data analysis for the early prediction of course-agnostic student performance," Computers & Education, Vol.163, 104108, 2021.

[5] C. R. Henrie, R. Bodily, R. Larsen, and C. R. Graham, "Exploring the potential of LMS log data as a proxy measure of student engagement," Journal of Computing in Higher Education, Vol.30, No.2 pp.344-362, 2018.

[6] D. Kim, Y. Park, M. Yoon, and I. H. Jo, "Toward evidence-based learning analyrics: Using proxy variables to improve asynchronous online discussion environments," Internet and Higher Education, Vol.30, pp.30-43, 2016.

[7] Y. Park, and I. H. Jo, "Using log variables in a learning management system to evaluate learning activity using the lens of activity theory," Assessment and Evaluation in Higher Education, Vol.42, No.4 pp.531-547, 2017.

[8] B. Rienties, L. Toetenel, and A. Bryan, ""Scaling up" learning design: impact of learning design activities on LMS behavior and performance," In Proceedings of the Fifth International Conference on Learning Analytics and Knowldege, pp.315-319, 2015.

[9] 성한울, 조일현, "온라인 학습 상황에서 행동 로그, 생리심리반응 및 시험불안을 통한 멀티모달(Multimodal) 학업성취 예측모형 개발," 교육공학연구, 제34권, 제2호, pp.287-308, 2018.

[10] 이해듬, "학습분석학 관점의 대학 이러닝 학습자 군집화와 학업성취도 관계 분석: 이러닝 학습 시·공간 데이터를 기반으로," 평생학습사회, 제14권, 제3호, pp.97-118, 2018.

[11] 이현진, "오토인코더에 기반한 딥러닝을 이용한 사이버대학교 학생의 학업 성취도 예측 분석 시스템 연구," 한국디지털콘텐츠학회 논문지, 제19권, 제6호, pp.1115-1121, 2018.

[12] 조일현, 김정현, "학습분석학을 활용한 e-러닝 학업성과 추정 모형의 통계적 유의성 확보 시점 규명," 교육공학연구, 제29권, 제2호, pp.285-306, 2013.

[13] 조헌국, "머신 러닝을 활용한 이러닝 학습 환경에서의 학습자 성취 예측 모형 탐색," 학습자중심교과교육연구, 제18권, 제21호, pp.553-572, 2018.

[14] R. Conjin, C. Snijders A. Kleingeld, and U. Matzat, "Predicting Student Performance from LMS Data: A Comparison of 17 Blended Courses Using Moodle LMS," IEEE Transactions on Learning Technologies, Vol.10, No.1, pp.17-29, 2016.

[15] Y. H. Hu, C. L. Lo, and S. P. Shih, "Developing early warning systems to predict students' online learning performance," Computers in Human Behavior, Vol.36, pp.469-478, 2014.

[16] J. G. Cromley, T. Perez, A. Kaplan, T. Dai, K. Mara, and M. J. Balsai, "Combined Cognitive-Motivational Modules Delivered Via an LMS Increase Undergraduate Biology Grades," Technology, Mind, and Behavior, Vol.1, No.2, 2020.

[17] 유진은, "기계학습: 대용량/패널자료와 학습분석학 자료 분석으로의 활용," 교육공학연구, 제35권, 제2호, pp.313-338, 2019.

[18] J. Heo, H. Lim, S. Yun, S. Ju, S. Park, and R. Lee, "Descriptive and Predictive Modeling of Student Achivement, Satisfaction, and Mental Health for Data-Driven Smart Connected Campus Life Service," In Proceedings of the 9th International Conference on Learning Analytics & Knowledge, pp.531-538, 2019.

[19] G. Van Rossum, and F. L. Drake, Python 3 Reference Manual, Scotts Valley, CA: CreateSpace, 2009.

[20] Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," In Advances in Neural Information Processing System 32, pp.8024-8035, 2019.

[21] M. Tietz et al., "skorch: A scikit-learn compatible neural network library that wraps PyTorch," July 2017. [Online]. Available:

http://skorch.readthedocs.io

[22] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, Vol.12, pp.2825-2830, 2011.

저 자 소 개

문 기 범(Kibum Moon) **정회원**

- 2015년 2월 : 고려대학교 심리학과 (문학사)
- 2018년 2월 : 고려대학교 임상 및 상담심리전공 (문학석사)
- 2019년 10월 ~ 현재 : 고려대학교 디지털정보처 데이터사이언티스트

<관심분야> : 인공지능, 빅데이터, 심리학

김 진 원(Jinwon Kim) **정회원**

- 2018년 2월 : 고려대학교 심리학과 (문학사)
- 2020년 8월 : 고려대학교 임상 및 상담심리전공 (문학석사)
- 2020년 10월 ~ 현재 : 고려대학교 디지털정보처 데이터사이언티스트

<관심분야> : 심리학, 인공지능, 데이터사이언스

이 진 숙(Jinsook Lee) **정회원**

- 2017년 2월 : 고려대학교 가정교육과 (가정학사)
- 2019년 2월 : 고려대학교 생활과학과 (이학석사)
- 2019년 7월 ~ 현재 : 고려대학교 디지털정보처 데이터사이언티스트

<관심분야> : 인공지능, 적응형학습, 교육공학