서비스 이탈 고객 예측 모델 분석 보고서

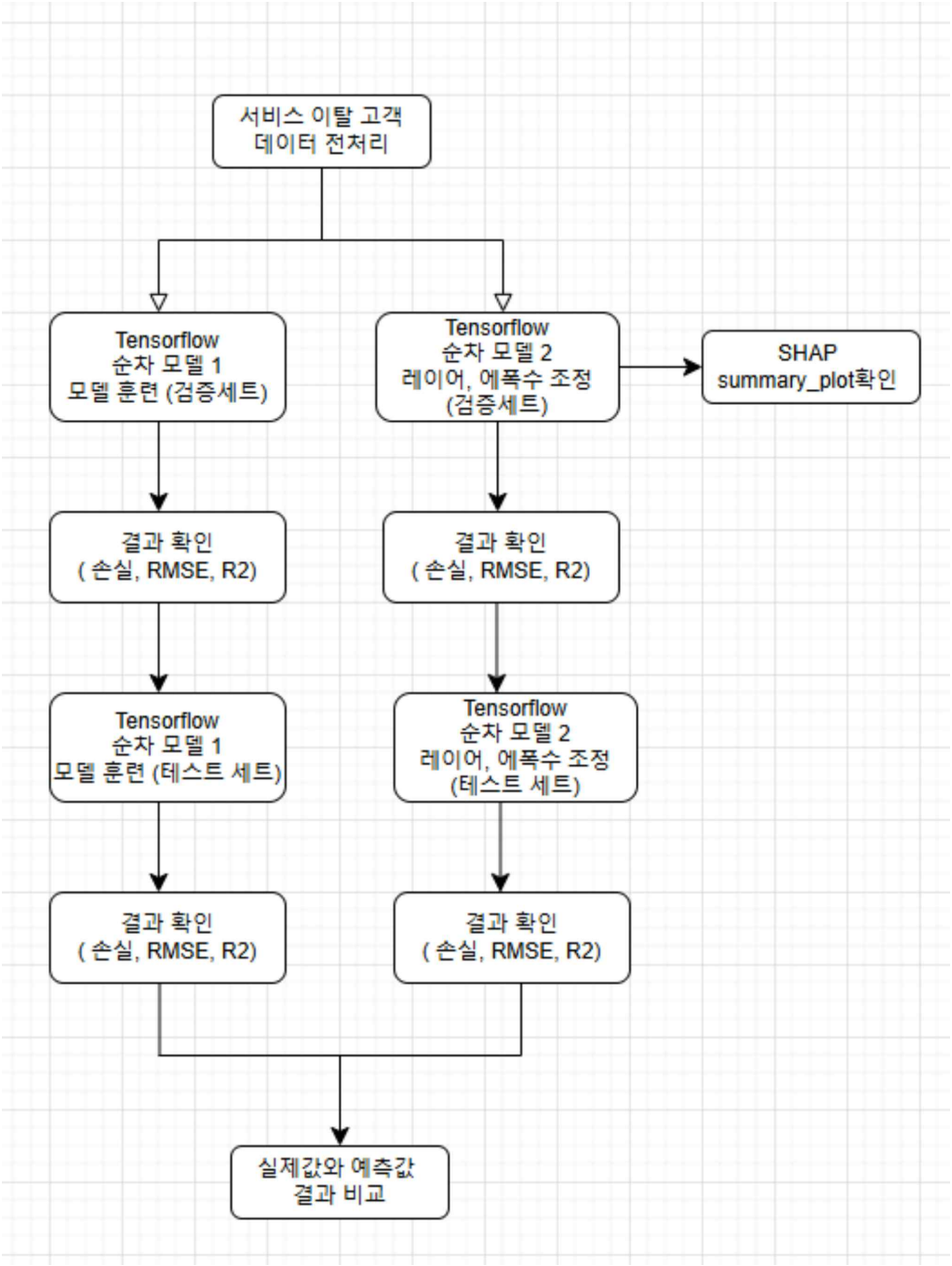
빅데이터 8기 이현희

요약 : 디지털 기기 및 플랫폼을 이용한 온라인 학습의 중요도가 높아짐에 따라, 온라인 학습자의 학습 형태 및 회원 형태를 분석하는 것이 중요해졌다. 이에 따라 서비스 이탈 고객 예측 모델을 생성하여 만료 일(change_date)를 예측하는 딥러닝 모델을 생성하였다. 순차모델을 사용하고 다양한 하이퍼파라미터를 조정하여 최종 모델을 구현하여, 손실 2.38, RMSE는 1.54, R^2 는 0.86의 결과를 보였다. 이 모델을 통해 실제 값과 예측 값을 비교하였더니, 실제값과 예측값이 유사하게 나왔음을 확인하여 이 모델을 토앫 온라인 학습 환경에서 다양한 교육관련 요소 및 특성들을 고려하여 서비스 이탈 고객 예측 모델을 생성하여 학습자의 이탈 시점을 예측하고자 한다. 이를 바탕으로 이탈 원인 및 추후 개선 방안까지 함께 도출할 수 있다.

Ⅰ. 서론

온라인 학습 시장은 학습자의 적극적인 온라인 학습 참여에 의존한다. 이에 따라, 학습자가 어떤 과목을 들으며, 어떤 학습 형태 및 학습 과정을 통해 학습을 완료 혹은 포기 하게 되는지 확인하는 것은 매우 중요하다. 학습자의 학습 과정에 영향을 미치는 요소는 외적인 요소와 내적인 요소 모두와 관련성이 있으며, 외적인 요소로는 수업 상황, 학습 환경, 외적 동기 부여 등이 있고, 내적인 요소로는 학습자의 동기 및 관심사 등이 있다. 이는 과목, 수업 방법, 수업 환경 등 여러 학습 요인과 관련성이 있다. 이에 따라, 본 분석에서는 학습자가 만료 회원이 되는 시기를 예측하는 것을 목적에 두고 있다.

다음은 본 분석의 진행 순서도이다.



[그림1. 순서도]

Ⅱ. 본론

1) 데이터 소개

① 데이터 형태 및 Column

본 데이터는 45개의 열과, 6476개의 행으로 이루어져있다. 여기서 45개의 칼럼은 다음 표와 같다.

Column	해석	개수	Column	해석	개수
userid	회원 고유 ID	30	time stamp	해당 메뉴 이동이 발생한 시간	개별
learning _seq	학습 순서	33	gender	성별	3
mcode	콘텐츠고유 ID	205	grade	학년	1
learning _action_ seq	학습 중 행동순서	147	member status	회원 상태	6
event_ type	이벤트 유형	3	memberstat us_change	월 중 회원 상태 변화	6
action	이벤트에 포함되는 행동 내용	10	day_00_ staus	00일 회원 상태	개별
object_ type	활동 대상	3	change_date	만료 회원이 된 날	12

[표1. 데이터 Column 소개]

② 학습자 회원 형태 소개

학습자의 회원 형태는 본 데이터에는 총 6가지의 형태로 나타나 있다. 회원 형태는 다음과 같다.

00	학습생 대기 - 무료	01	학습생 (준) - 무료
02	학습생 (일반) - 무료	11	학습생 (정) - 유료
44	만료 회원 - 만료	55	정회원_이월 - 유료

[표2. 학습자 회원 형태 소개]

2) 데이터 전처리

① 원핫 인코딩

원-핫 인코딩(One-Hot Encoding)을 이용하여 해당 데이터의 열 중 수치형 데이터를 범주형 데이터로 변환해주었다.. 구체적으로 'mcode', 'event_type', 'action', 'object_type', 'gender', 'grade' 6개의 열을 원핫 인코딩을 이용하여 범주형 데이터로 변환해주었다.

Column	원래 Column 개수	늘어난 Column 개수
전체	45	264
mcode	1	205
event_type	1	3
action	1	10
object_type	1	3

gender	1	3
grade	1	1

[표3. 원핫 인코딩 결과]

② 사용한 칼럼만 따로 정리 (active_data 에 정리)

active_data	
target	change_data
input	learning_seq
	learning_action_seq
	mcode (원-핫인코딩 205개)
	event_type (원-핫인코딩 3개)
	action (원-핫인코딩 10개)
	object_type (원-핫인코딩 3개)
	gender (원-핫인코딩 3개)

[표4. 사용할 데이터를 active_data에 저장]

3) Train, Val, Test로 데이터 나누기 및 정규화

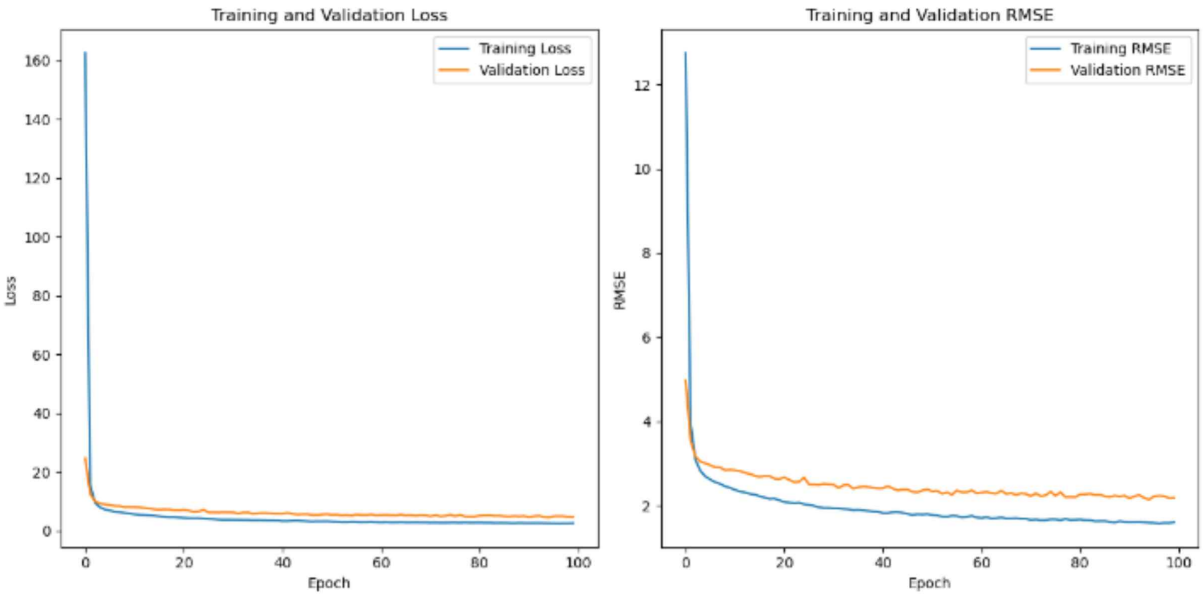
6476 x 226 의 형태를 지닌 데이터를 검증세트와 테스트 세트로 나누었다. 전체 데이터에서 8:2로 훈련과 테스트 세트를 나누고, 훈련 세트를 다시 8:2로 나누어 훈련 세트와 검증 세트로 나누었다 훈련, 검증, 테스트 세트의 크기는 다음과 같이 정리할 수 있다.

세트	크기
훈련 세트 크기	(4144, 226)
검증 세트 크기	(1036, 226)
테스트 세트 크기	(1296, 226)

[표4. train, val, test 세트 크기]

4) 모델 구축 및 훈련

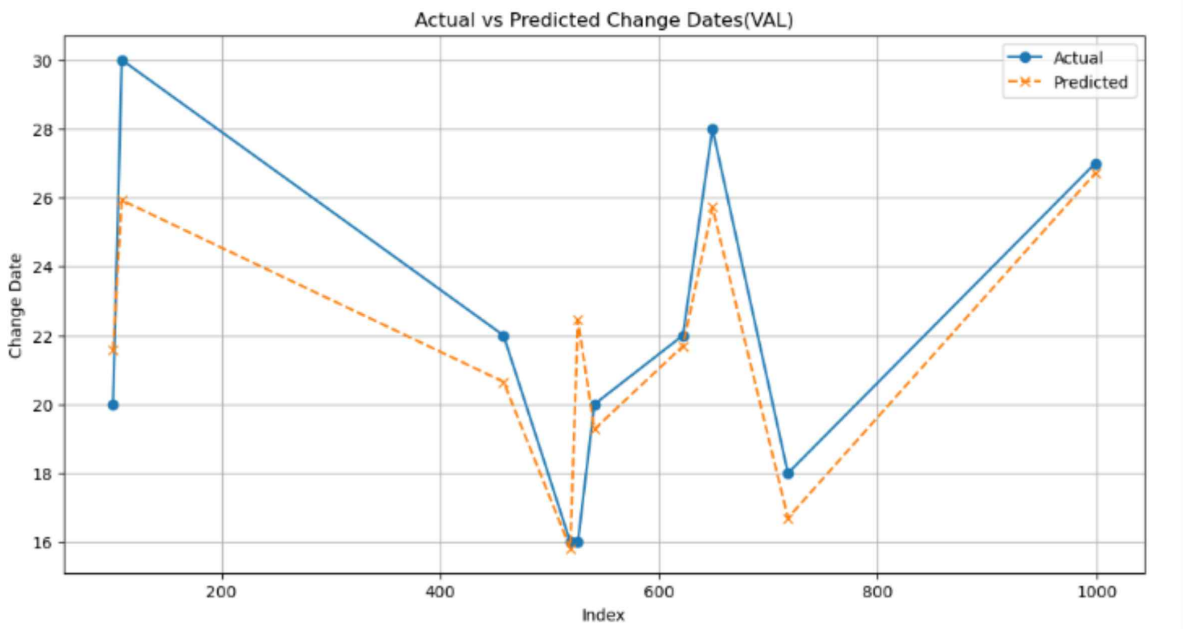
이에 따라, input은 change_date를 제외한 225개로 output은 change_date 1개로 설정하였고, 훈련세트, 테스트 세트, 검증 세트로 분할하여 모델을 훈련하였다. 훈련 전 X_train, X_val, X_test를 정규화하였다. Tensorflow에서 순차모델을 사용하여 모델을 훈련하였다. Dense 레이어는 총 3개로 구성되었으며, 최적화 함수로는 Adam optimizer를 손실함수로는 Mean Squared Error (MSE) 사용하였고, 평가 지표로는 Root Mean Squared Error (RMSE) 사용하였다. Epochs 수를 100으로 설정하여 모델을 훈련시킨 결과는 다음과 같다.



[그래프1. 모델 구축 및 훈련 검증세트 결과]

위 그래프에서 보이는 것 처럼 검증세트의 손실은 4.756, 검증 세트의 RMSE는 2.180로 R^2 는 0.7415의 점수를 보였다.

다음은 훈련된 모델을 이용하여 10개의 행을 랜덤하게 선택하여 검증 세트에서 실제와 예측 값을 비교한 그래프이다.



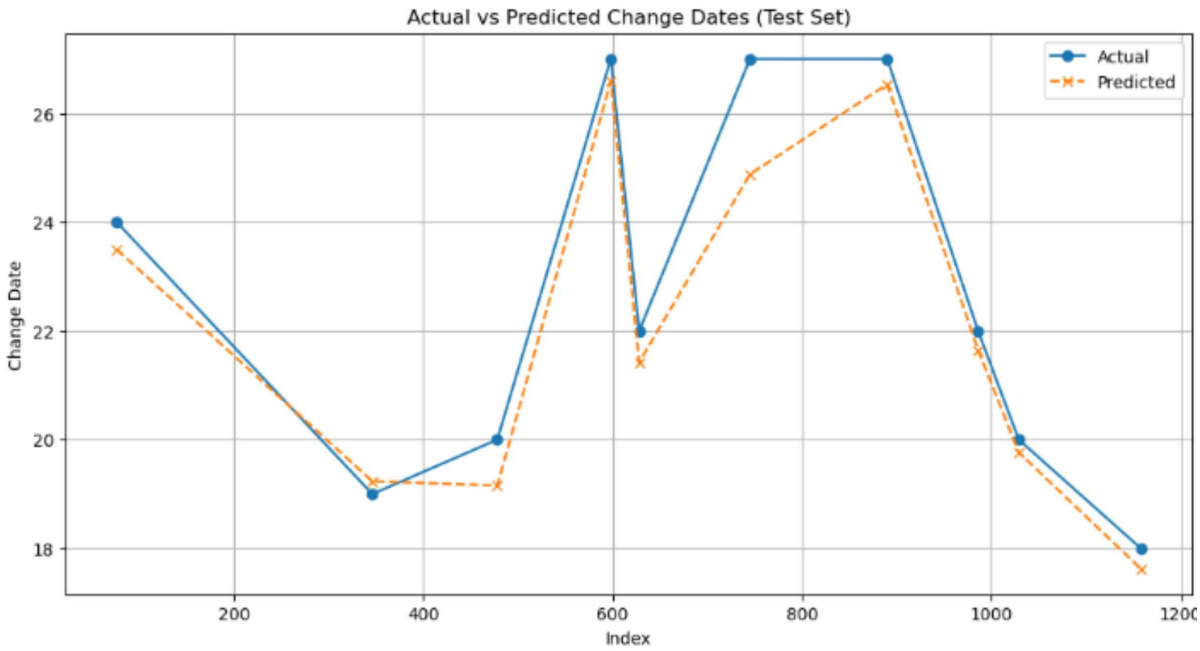
[그래프2. 모델 구축 및 훈련 검증 세트의 실제와 예측 비교 결과]

실제 값과 예측값이 유사한 행도 있지만, 4일 정도 차이가 나는 예측도 있다는 것을 보아, 모델

성능 개선이 필요하다고 판단하였다.

검증 세트 결과를 확인 한 후 테스트 결과도 함께 확인하였다. 테스트 세트의 손실은 3.96이었으며, 테스트 세트의 RMSE는 1.99, R^2는 0.77의 점수를 보였다.

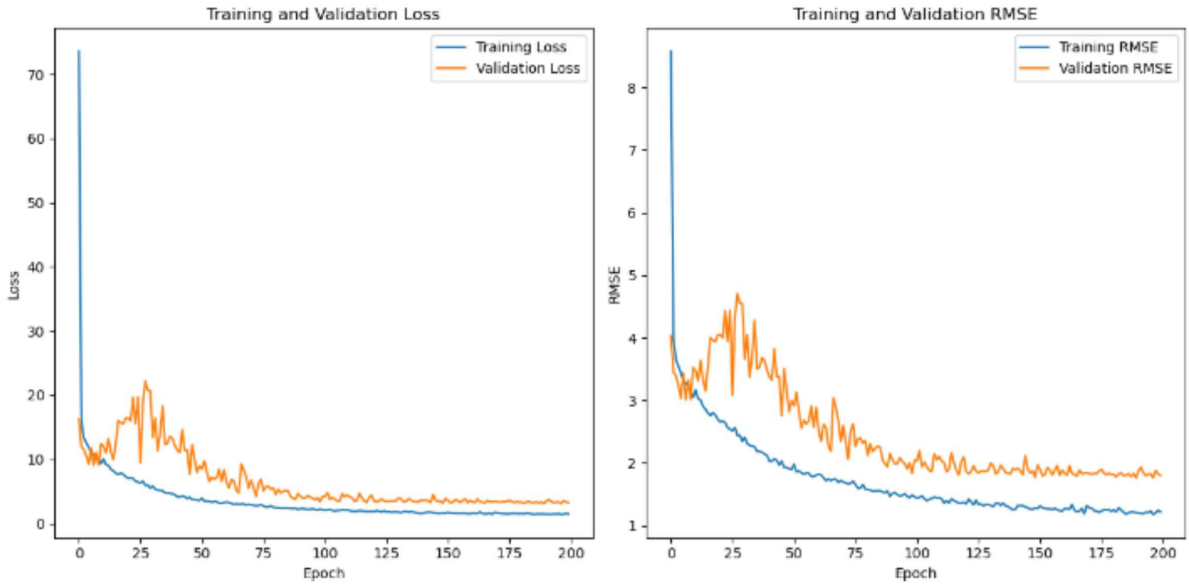
다음은 훈련된 모델을 이용하여 따라 10개의 행을 랜덤하게 선택하여 검증 세트에서 실제와 예측 값을 비교한 그래프이다.



[그래프3. 모델 구축 및 훈련 테스트 세트의 실제와 예측 비교 결과]

4) 모델 수정 및 훈련

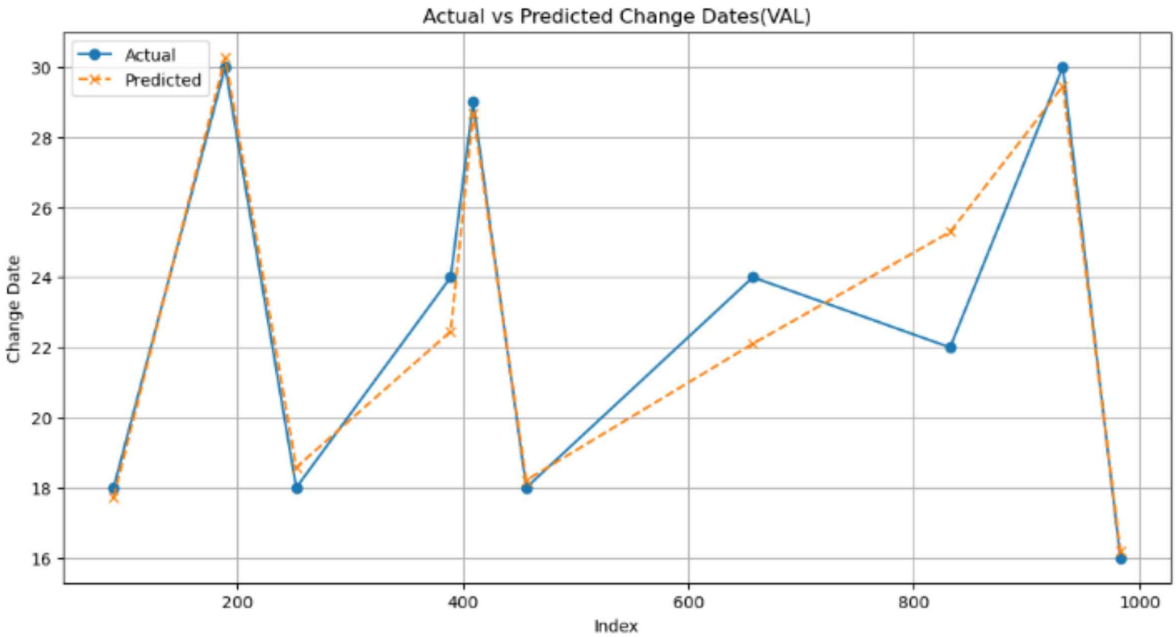
첫 번째 모델에서의 손실을 줄이기 위해 모델에 추가적으로 레이어, 에폭수를 조정하였다. 동일하게 순차 모델을 사용했으며 Dense 레이어를 3개에서 5개로 늘리고, 과적합을 방지하기 위해 dropout 레이어도 2개 추가하였다. 최적화 함수, 손실 함수, 평가 지표는 모델 훈련 1과 동일하게 사용하였다. 에폭수는 200으로 조정하여 모델을 훈련시킨 결과는 다음과 같다.



[그래프4. 모델 수정 및 훈련 검증세트 결과]

위 그래프에서 보이는 것 처럼 검증세트의 손실은 3.25 검증 세트의 RMSE는 1.80로 R^2 는 0.82 의 점수를 보였다.

다음은 훈련된 모델을 이용하여 따라 10개의 행을 랜덤하게 선택하여 검증 세트에서 실제와 예측 값을 비교한 그래프이다.



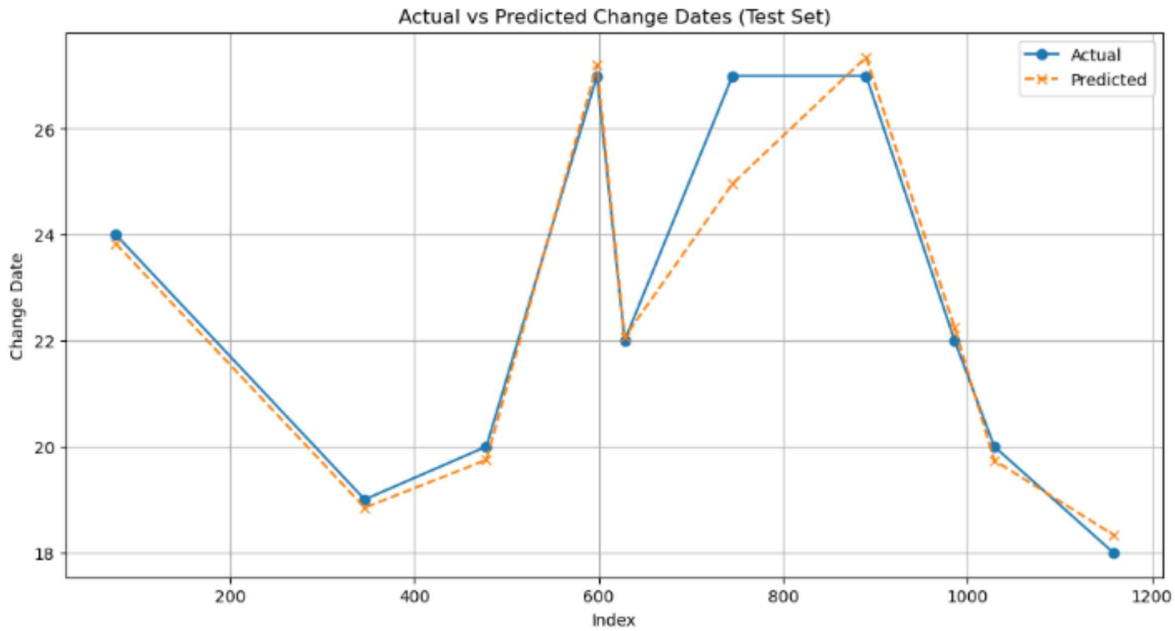
[그래프5. 모델 수정 및 훈련 검증 세트의 실제와 예측 비교 결과]

위 그래프를 처음 구축한 모델에 비해 레이어와 dropout을 추가한 모델이 실제값과 예측값이 비

슷하게 나왔음을 알 수 있다.

검증 세트 결과를 확인 한 후 테스트 결과도 함께 확인하였다. 테스트 세트의 손실은 2.38이었으며, 테스트 세트의 RMS는 1.54, R^2는 0.86의 점수를 보였다.

다음은 훈련된 모델을 이용하여 따라 10개의 행을 랜덤하게 선택하여 검증 세트에서 실제와 예측값을 비교한 그래프이다.

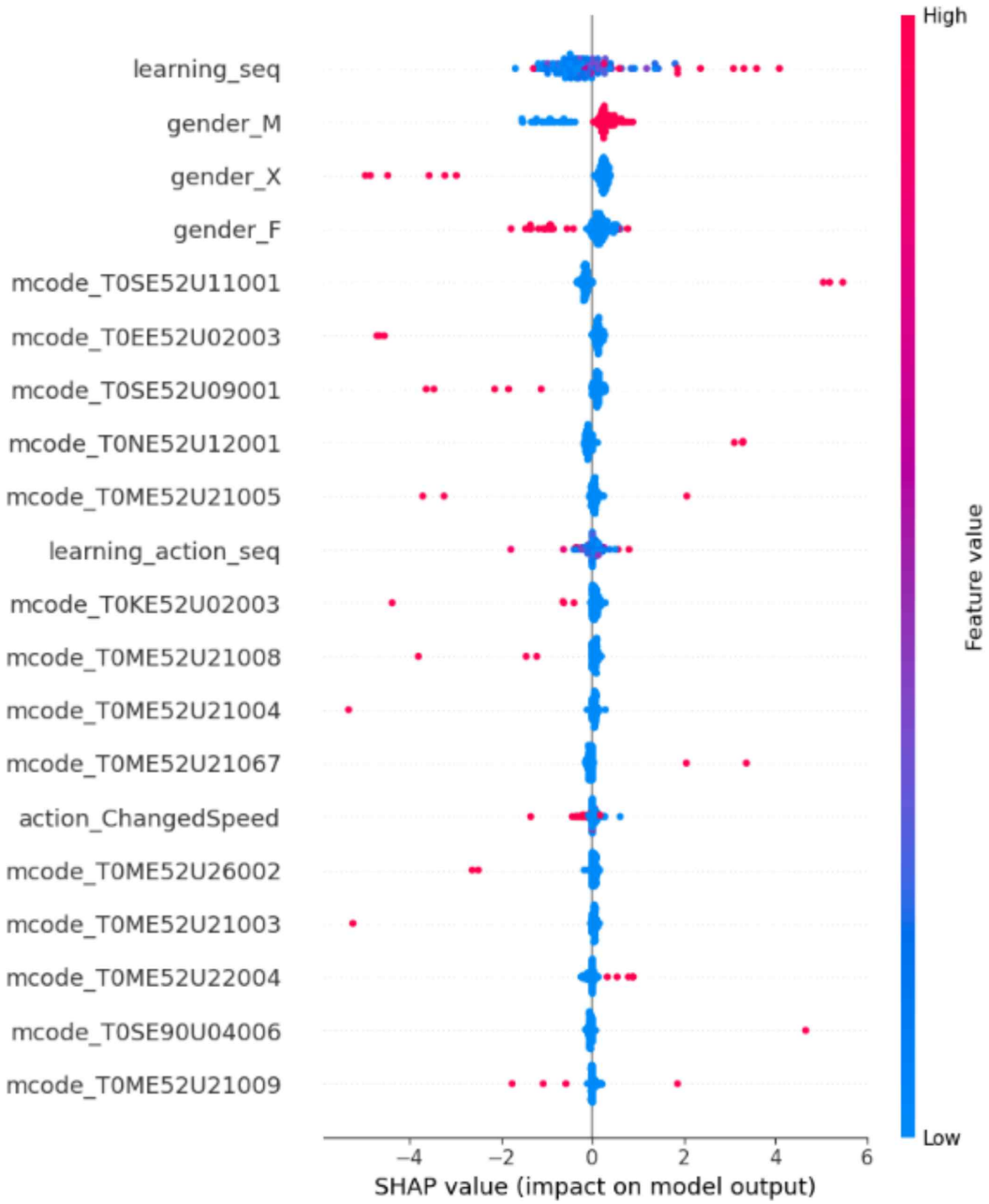


[그래프6. 모델 수정 및 훈련 테스트 세트의 실제와 예측 비교 결과]

위 그래프를 참고하여 테스트 세트에서 실제와 예측값이 가깝게 나타났음을 알 수 있다.

5) SHAP 사용하여 특성 중요도 확인

추가적으로 모델의 성능을 높이기 위해 SHAP를 사용하여 Feature간의 특성 중요도를 확인하였다. Y축 (세로축)에는 각 피처(Feature)들이 나열되어 있으며, 총 226개의 Feature가 있다. X축 (가로축): SHAP 값으로, 각 피처가 모델의 예측 결과에 미치는 영향을 나타낸다. SHAP 값이 양수이면 해당 피처가 예측값을 증가시키고, 음수면 예측값을 감소시킨다. 해당 모델 훈련과 관련된 SHAP summary_plot은 다음과 같다.



[그래프7. SHAP 그래프]

해당 그래프를 보면 learning_seq와 gender의 칼럼들이 가장 모델의 예측 결과에 영향을 미침을 알 수 있. 그 이 반면 mcode의 칼럼들은 모델의 예측값에 상대적으로 작은 영향을 미침을 알 수 있다.

Ⅲ. 결론

모델 훈련 및 수정을 통해 예측값이 실제값에 가까워질 수 있도록 하였으며, 모델 성능을 높였다. test 세트를 통해 훈련시킨 모델로 회원의 만료 날짜를 예측하였으며, 다음은 모델 수정 및 훈련

한 최종 모델을 바탕으로 test 세트에서 랜덤하게 5개의 행을 선택하여 change_date를 예측한 날짜와 실제 날짜를 정리한 표이다.

행	Actual	Predicted
2095	18	18.06
2415	16	15.96
3045	20	20.06
351	27	26.69
5494	24	23.69

[표6. 실제와 예측 값 비교]

위 표를 확인해보면 소수점을 반올림 하여 정수 값으로 보았을때, 예측값과 실제 값이 동일하게 나왔음을 알 수 있다. 이를 통해 모델 성능을 점검 및 확인하였다.

Ⅳ. 토론

이 모델을 통해 온라인 학습 환경에서 학습자, 즉 회원의 이탈 시점을 예측할 수 있다. 이 연장선에서 온라인 학습 관리 측면에서 학습자를 만료 회원으로 변화시키는 요인 등을 파악하여 요소들을 개선할 수 있다. 학습자의 수준과 관심 분야를 고려하여 학습 동기를 높여줄 수 있는 콘텐츠 제작, 집중할 수 있도록 즉각적 피드백 제공 등 다양한 요소에 변화를 주어 학습자의 학습 의욕을 높여갈 수 있다.

구체적으로, 학습자 맞춤형 학습 경로 및 AI 튜터를 이용한 개별화 피드백를 제공할 수 있다. 학습 속도, 학습 과정 중의 멈춤, 진단평가, 형성평가 등을 통해 학습자의 학습 성향과 출발점 및 학습 역량을 기록하여 학습자에게 맞춤형 학습 경로를 제공할 수 있다. 이를 위해 학습 과정 중 AI 튜터 기능을 이용할 수 있다. AI 튜터는 학습자의 성과와 어려움을 실시간으로 분석하여, 각 학습자의 요구와 수준에 맞춘 맞춤형 조언을 제공함으로써 학습의 효율성을 높이고, 학습자가 자신의 학습 목표를 명확히 설정하고 달성할 수 있도록 돕는다. AI 튜터 기능을 이용하여 개별화 피드백을 제공해주며 학습자의 학습 의욕 및 학습 과정의 개선에 도움을 줄 수 있다. 이를 통해, 온라인 학습 환경에서 자기주도 학습을 기르는 것뿐만 아니라 적절한 스캐폴딩으로 학습자 스스로 학습 수준을 향상시킬 수 있도록 지원할 수 있다.

결론적으로, 맞춤형 학습 경로와 AI 기반 개별화 피드백을 통해 학습자는 자신의 학습 여정을 보다 효과적으로 관리하고, 자기주도 학습을 통해 더 높은 학습 성과를 달성할 수 있다. 이는 온라인 학습 환경의 품질을 높이고, 학습자의 지속적인 참여와 학습 목표 달성을 지원하는 중요한 방법이 될 것이라 기대한다.

<사용한 패키지 버전>

1. matplotlib 3.9.1
2. numpy 1.26.4
3. python 3.12.3
4. pandas 2.2.2
5. shap 0.46.0
6. seaborn 0.13.2
7. tensorflow 2.17.0