밀크 T 학년별 이탈률 분석 보고서

: 전처리에 따른 성능 평가

프로젝트 기반 빅데이터 서비스 개발자 양성 과정 (4기)
이형석

결론

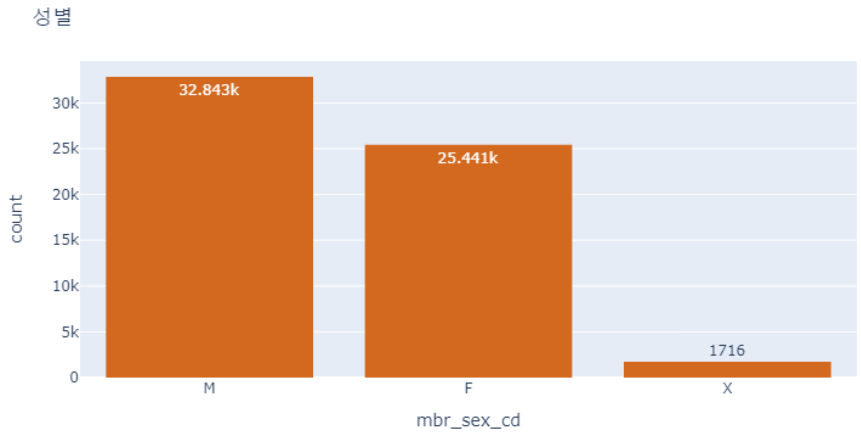
이 프로젝트는 전처리 방법에 따라 모델 학습 성능은 어떻게 달라질지 실험하는 것에 목적이 있다. 기존의 데이터와 전처리 후 데이터는 큰 차이를 보이지는 않았지만, 많게는 4%정도의 성능 차이를 보였다. 실험을 위해 같은 파라미터를 가진 모델에 원본 데이터와 전처리 된 데이터를 학습시켜 성능을 비교해 봤으며, 미이탈률이 높은 비율을 차지하고 있어 accuracy 는 대부분 99%를 보였다. 각 학년별로 성능을 비교한 결과 2 학년과 4 학년을 제외한 1,3,5,6 학년에서 성능 변화가 있었으며 1 학년의 precision 지표의 변화가 4%로 가장 높은 변화를 보여줬다.

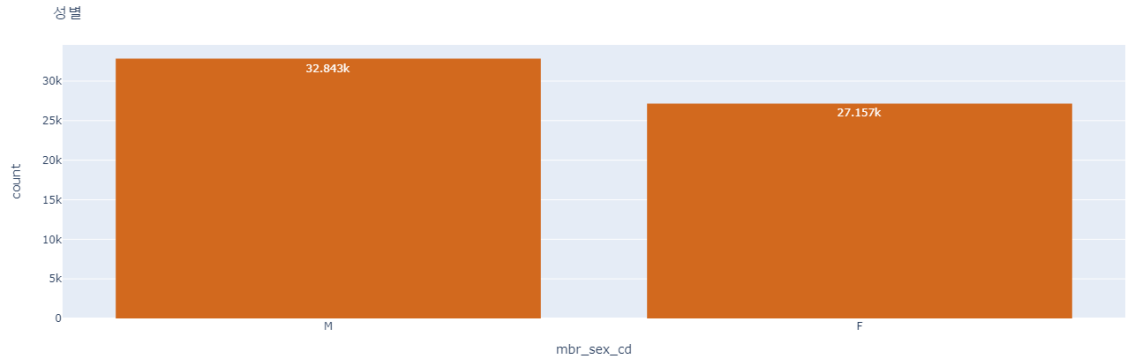
실험

전처리 및 시각화

- 'mbr_sex_cd'

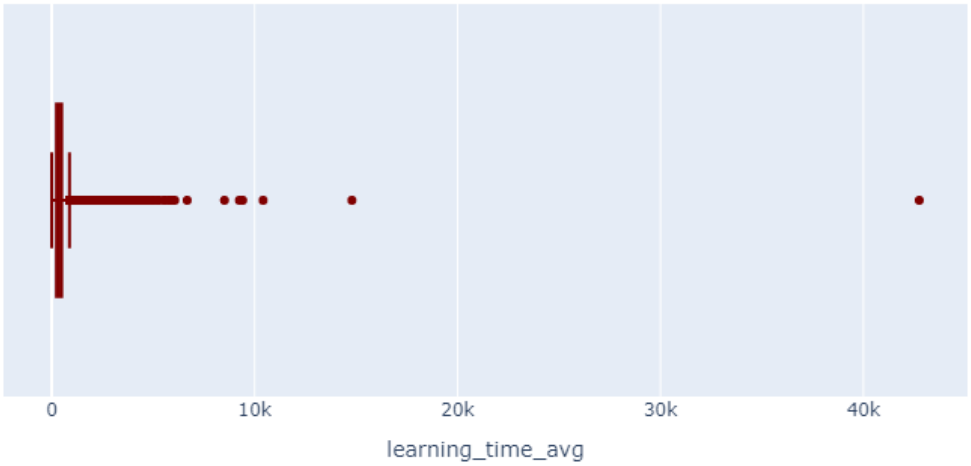
mbr_sex_cd 컬럼에는 'X'값으로 어떤 성별정보도 나타내지 않는 데이터가 있다. 성별에 따른 이탈률도 관련이 있을 것이라 생각해 'X'값을 제외한 성별 데이터의 비율을 확인하고 그 비율대로 'X' 데이터를 대체했다.

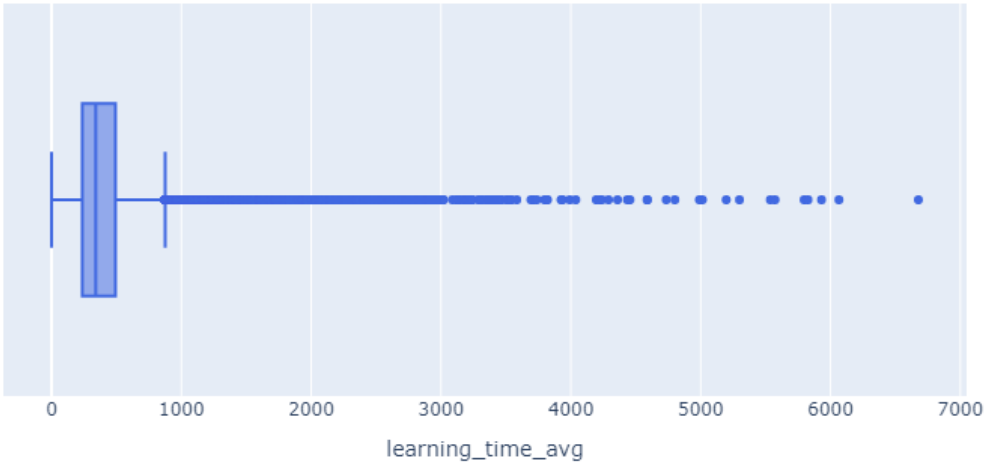




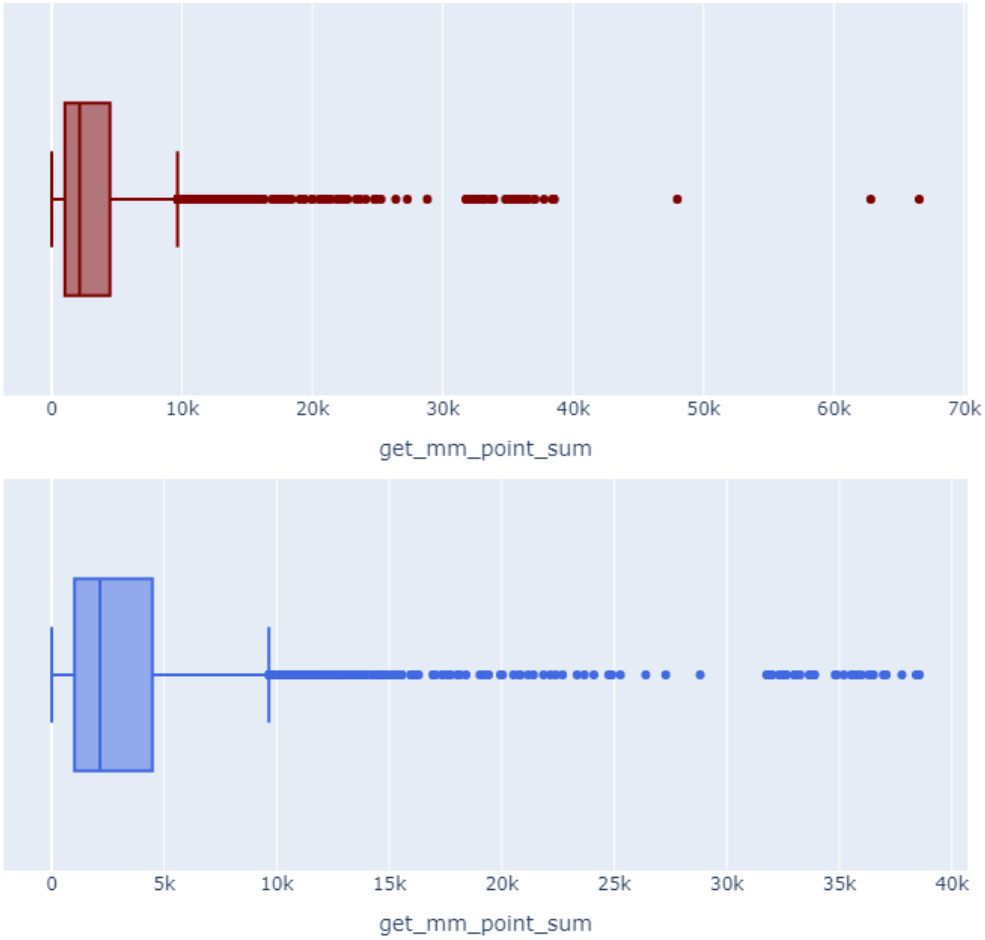
- 이상치 제거

연속형 변수를 시각화 한 결과 get_mm_point_sum, learning_time_avg, media_action_cnt_sum, non_video_veiwed_cnt_sum 4 개 컬럼에 이상치가 있다고 판단되어 사분위수를 활용해 이상치 찾아냈다. 이상치를 갖고 있지만 그 외 컬럼에 유의미한 데이터가 있을 수 있는 가능성이 있기 때문에 제거하지 않고 평균값으로 대체했다.

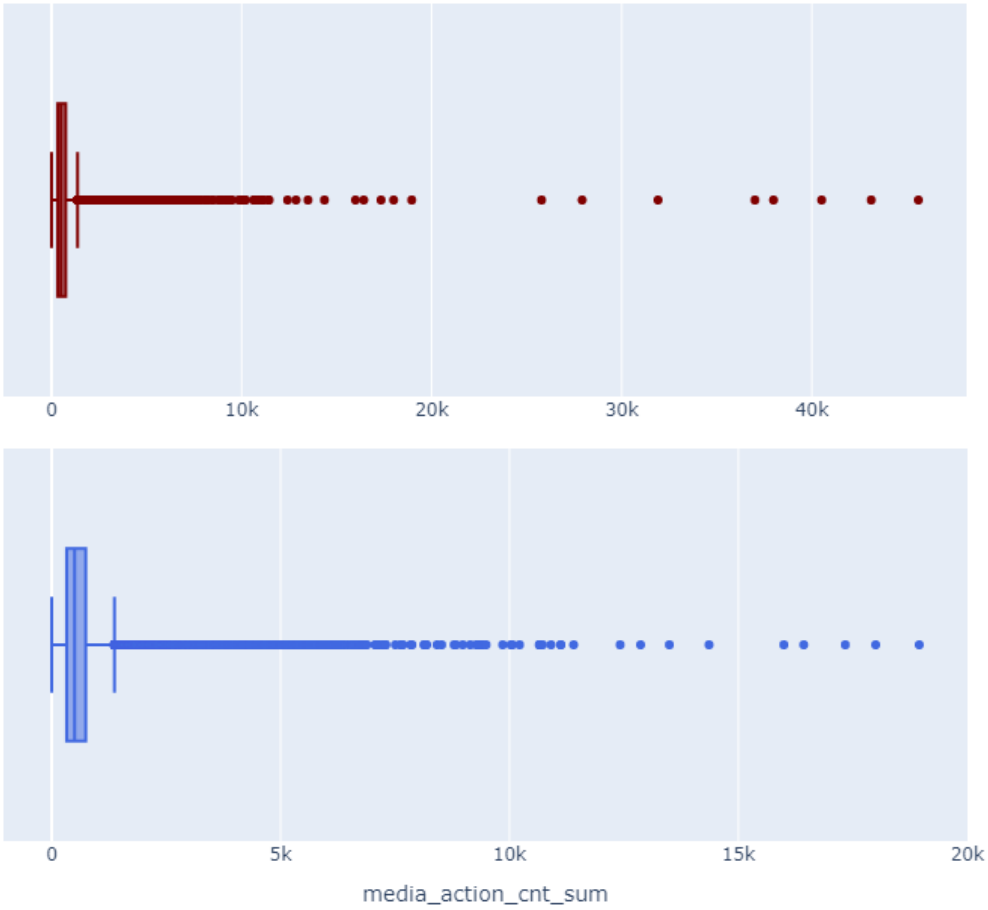




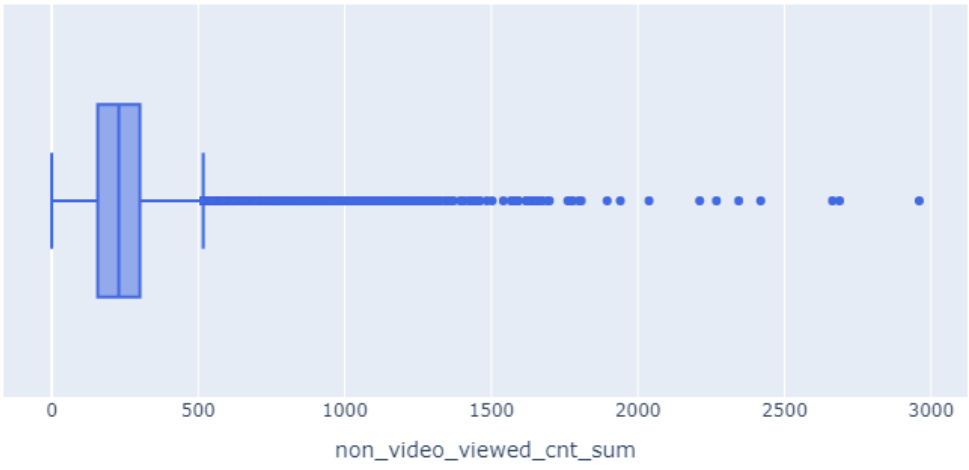
당월 학습 시간 평균을 나타내는 learning_time_avg 컬럼은 약 7000 을 기준으로 총 6 개 데이터를 평균값으로 대체했다.

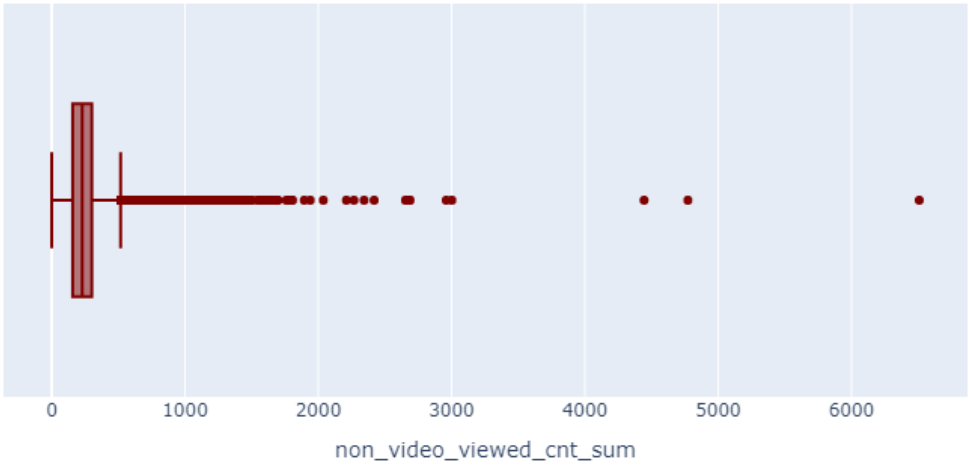


당월 획득 점수 평균를 나타내는 get_mm_point_sum 컬럼은 약 40,000 을 기준으로 총 3 개의 값을 평균값으로 대체했다.



미디어 콘텐츠 내 동영상 행동 횟수 (총합)을 나타내는 media_action_cnt_sum 컬럼은 20,000 을 기준으로 총 8 개의 값을 평균값으로 대체했다.



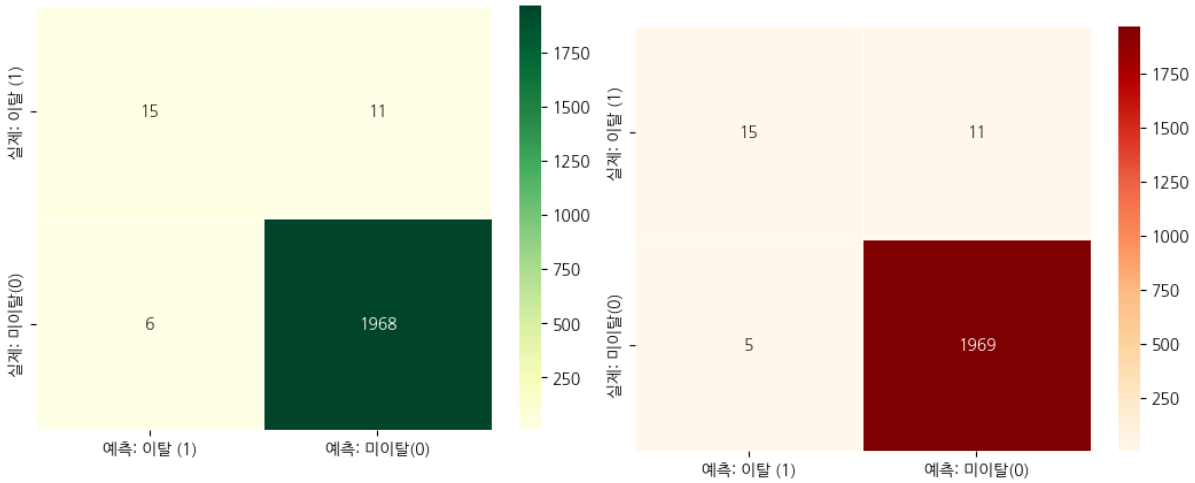


미디어 콘텐츠 미시청 행동 횟수 (총합)을 나타내는 non_video_viewed_cnt_sum 컬럼은 3000 을 기준으로 총 3 개의 데이터를 평균값으로 대체했다.

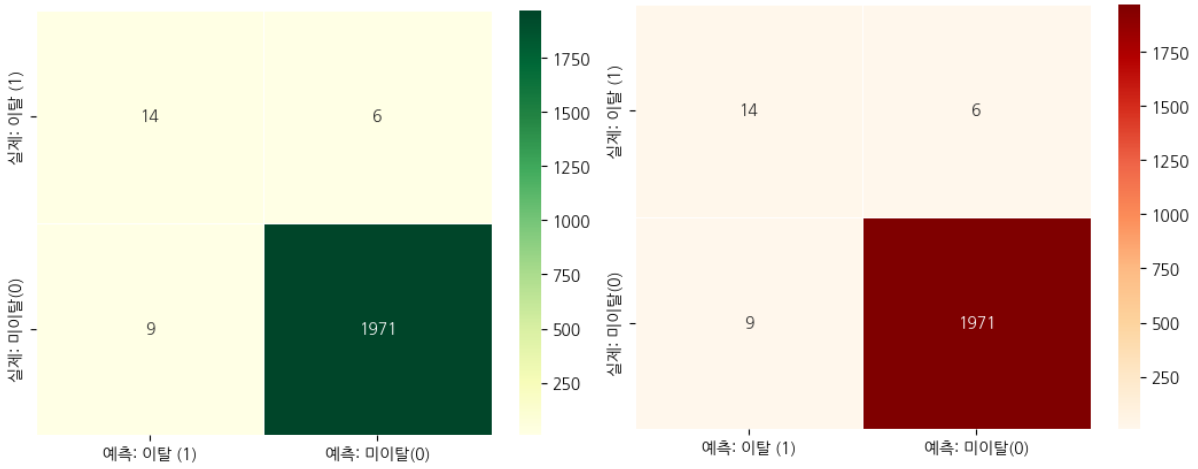
모델링

- 학습 모델은 LightGBM 을 활용했으며 파라미터는 default 값으로 설정했다.

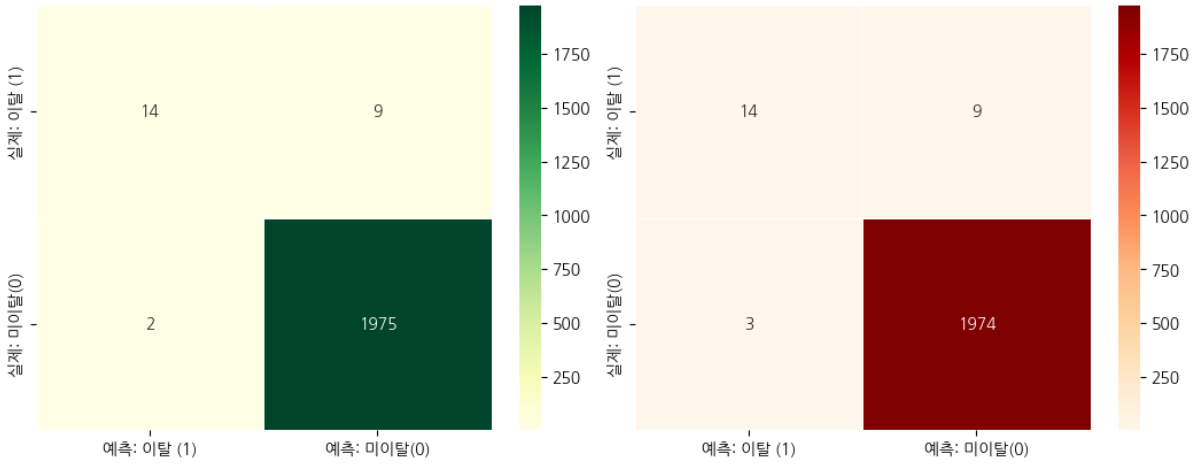
1 학년	Raw Data	Preprocess Data
Accuracy	0.99	0.99
Precision	0.71	0.75
Recall	0.58	0.58
F1-Score	0.64	0.65



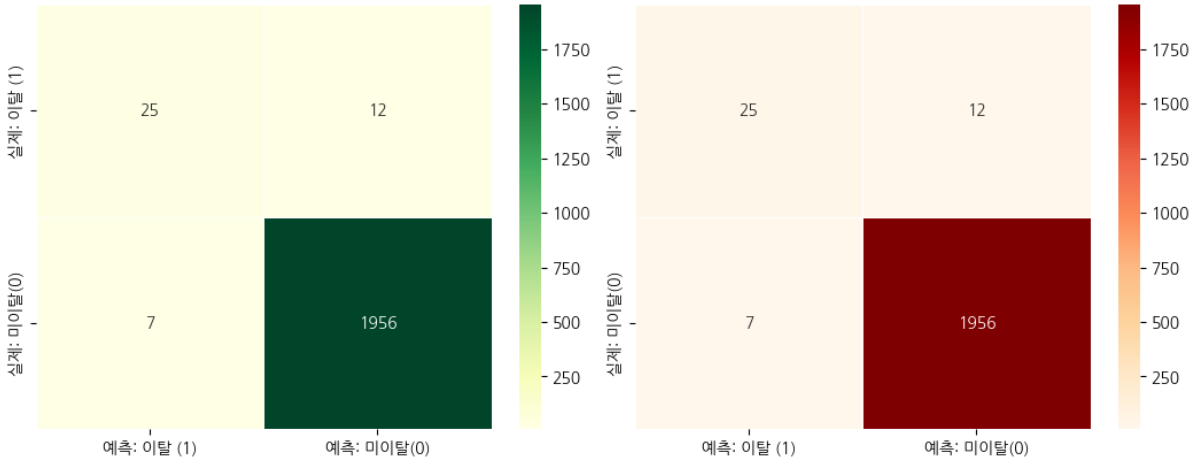
2 학년	Raw Data	Preprocess Data
Accuracy	0.99	0.99
Precision	0.61	0.61
Recall	0.7	0.7
F1-Score	0.65	0.65



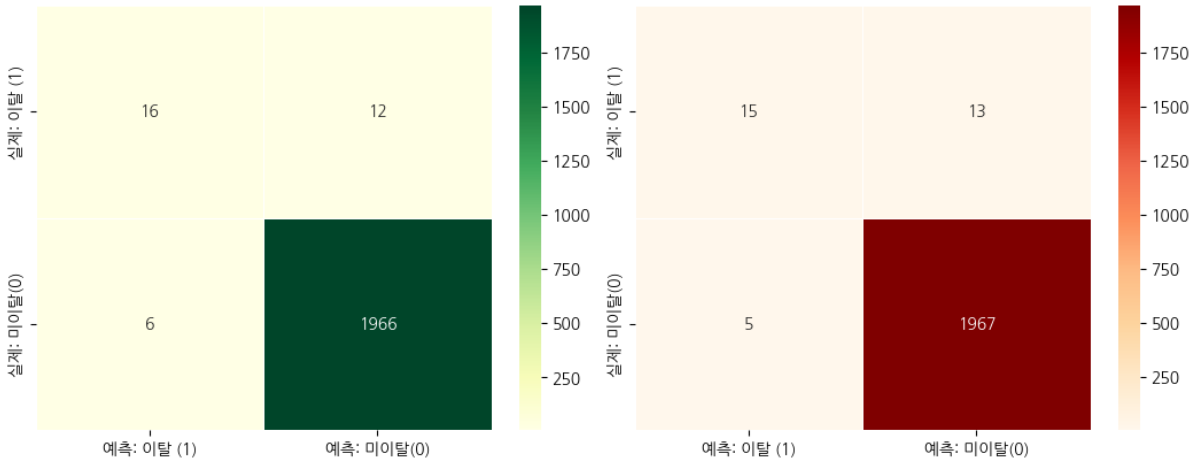
3 학년	Raw Data	Preprocess Data
Accuracy	0.99	0.99
Precision	0.88	0.82
Recall	0.61	0.61
F1-Score	0.72	0.7



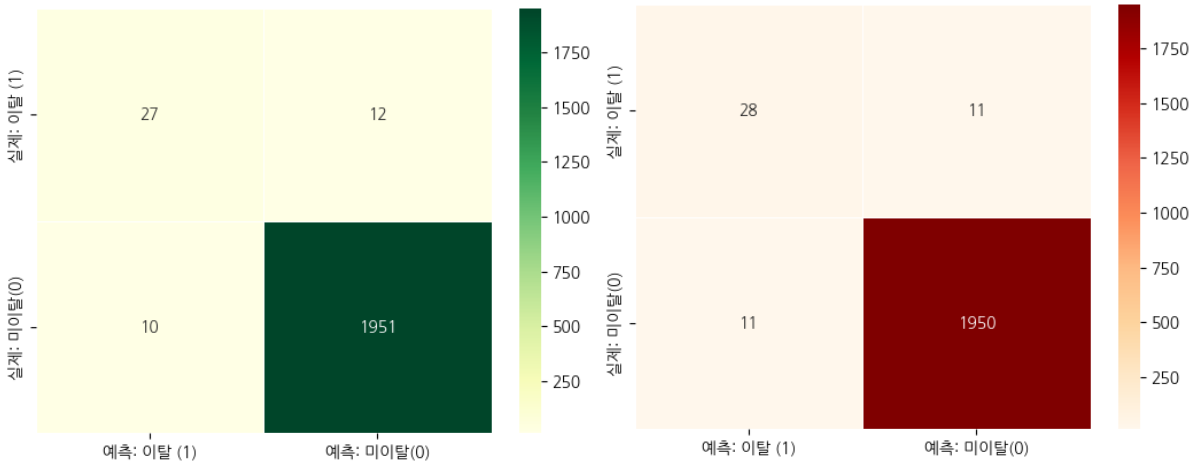
4 학년	Raw Data	Preprocess Data
Accuracy	0.99	0.99
Precision	0.78	0.78
Recall	0.68	0.68
F1-Score	0.72	0.72



5 학년	Raw Data	Preprocess Data
Accuracy	0.99	0.99
Precision	0.72	0.75
Recall	0.57	0.54
F1-Score	0.64	0.63



6 학년	Raw Data	Preprocess Data
Accuracy	0.99	0.99
Precision	0.73	0.72
Recall	0.69	0.72
F1-Score	0.71	0.72



작업 코드

https://github.com/lhshs/Genia/blob/main/Study/2308_Caliper2/day1/report.ipynb