**밀크T 학년별 이탈률 분석 보고서**

**:전처리에 따른 성능 평가**

프로젝트 기반 빅데이터 서비스 개발자 양성 과정 (4기)  
이형석

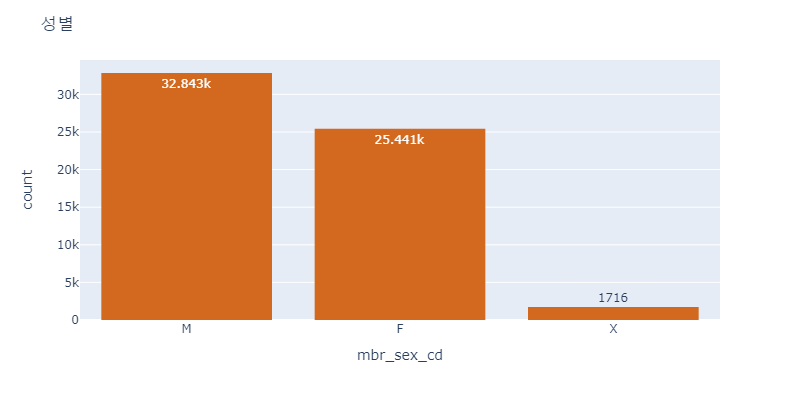
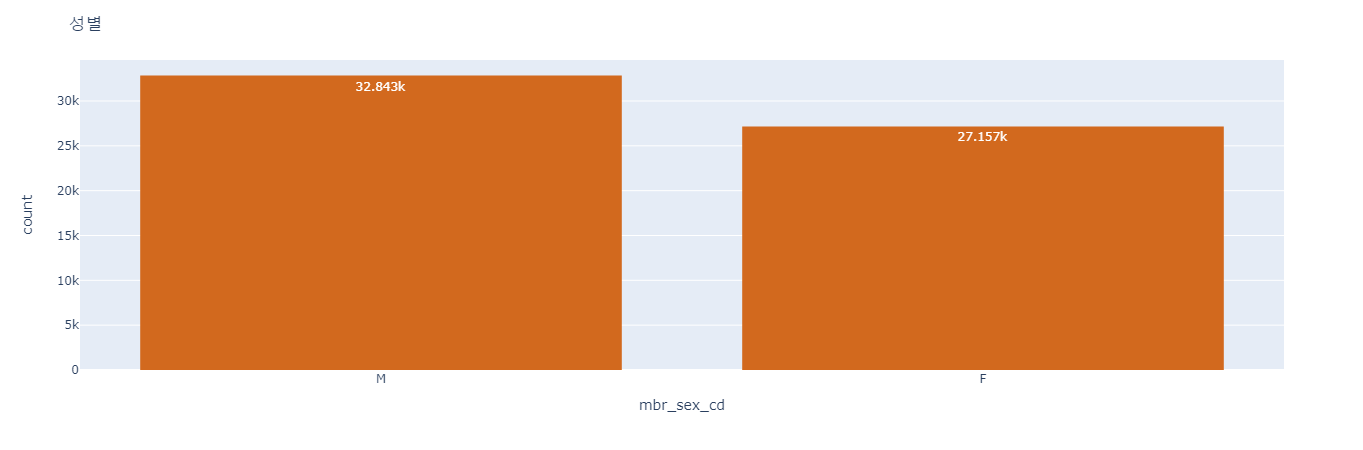
**요약**

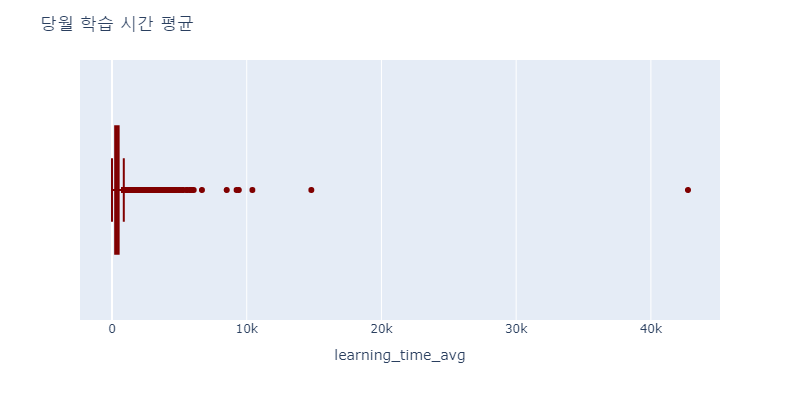
**개요**

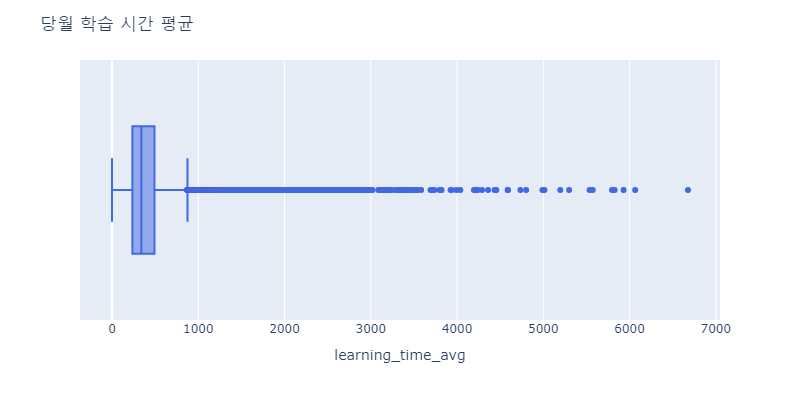
수집한 데이터의 학년별 이탈률에 영향을 끼치는 중요 변수를 찾아보고 Raw 데이터와 전처리한 데이터를 중심으로 모델의 성능 향상을 실험했다.

**실험**

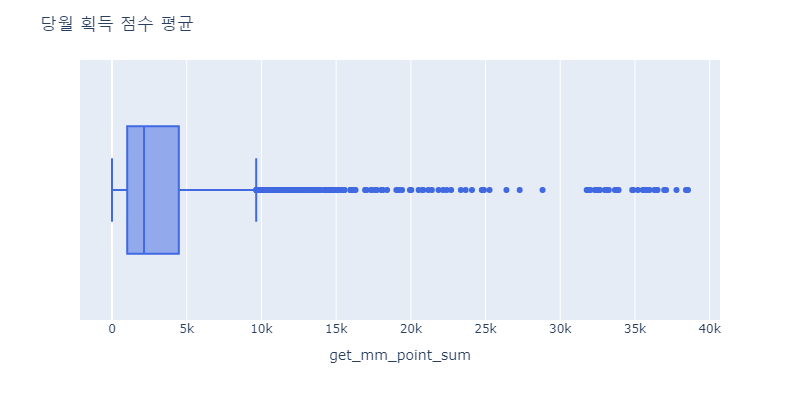
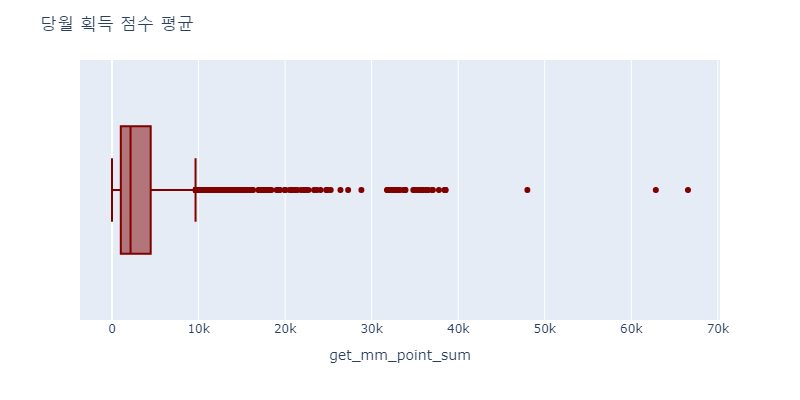
**전처리 및 시각화**

- ‘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.   
****

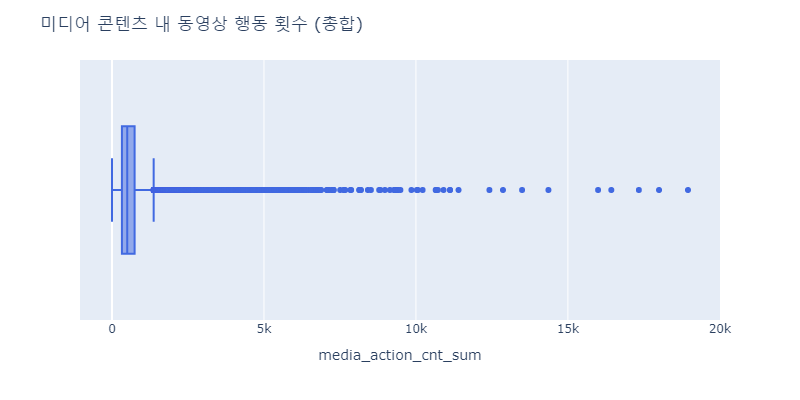
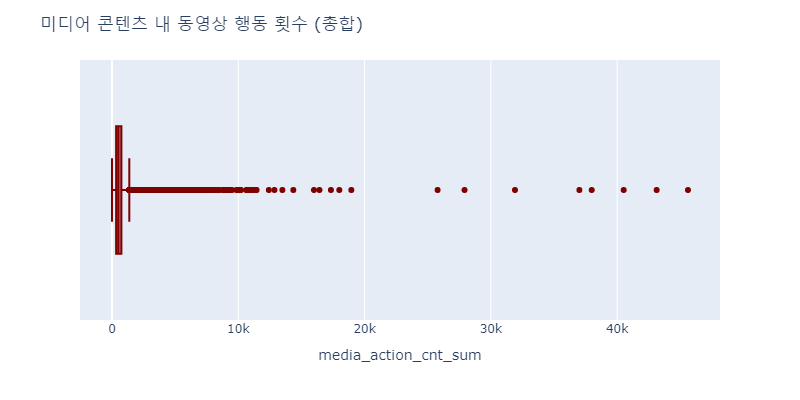
- 이상치 제거  
연속형 변수를 시각화 한 결과 get\_mm\_point\_sum, learning\_time\_avg, media\_action\_cnt\_sum, non\_video\_veiwed\_cnt\_sum 4개 컬럼에 이상치가 있다고 판단되어 사분위수를 활용해 이상치 찾아냈다. 이상치를 갖고 있지만 그 외 컬럼에 유의미한 데이터가 있을 수 있는 가능성이 있기 때문에 제거하지 않고 평균값으로 대체했다.



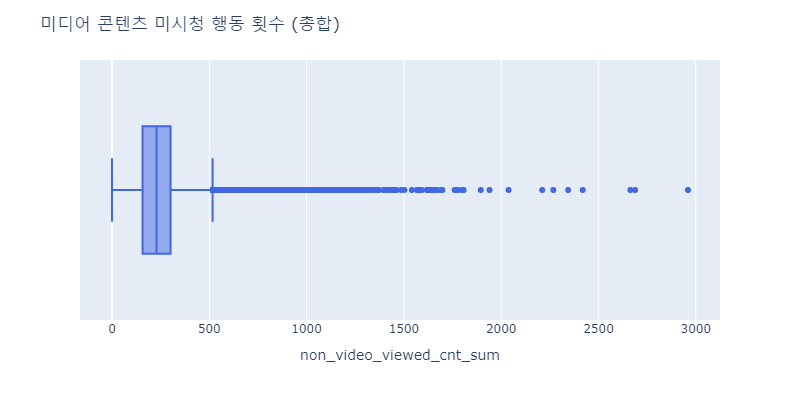
당월 학습 시간 평균을 나타내는 learning\_time\_avg 컬럼은 약 7000을 기준으로 총 6개 데이터를 평균값으로 대체했다.

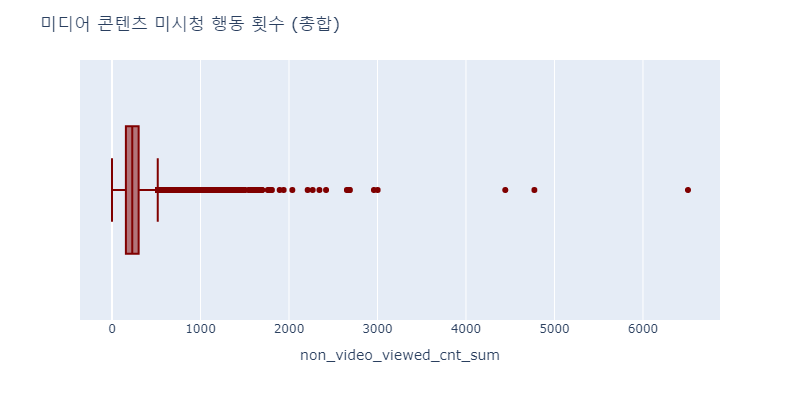


당월 획득 점수 평균를 나타내는 get\_mm\_point\_sum 컬럼은 약 40,000을 기준으로 총3개의 값을 평균값으로 대체했다.



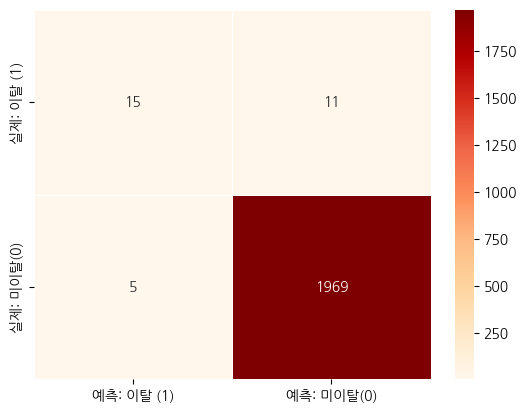
미디어 콘텐츠 내 동영상 행동 횟수 (총합)을 나타내는 media\_action\_cnt\_sum 컬럼은 20,000을 기준으로 총 8개의 값을 평균값으로 대체했다.



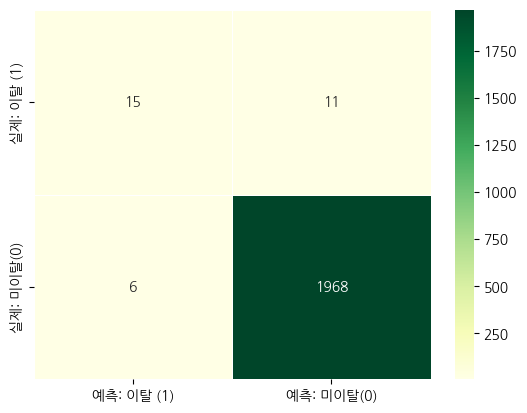


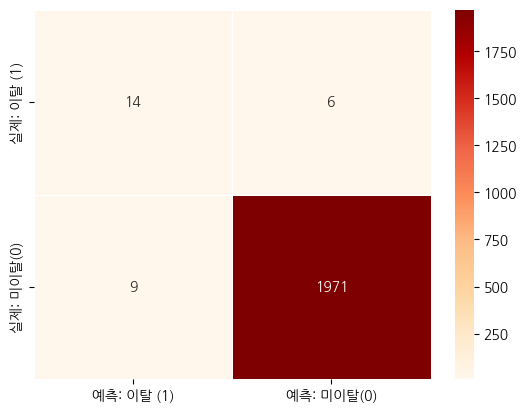
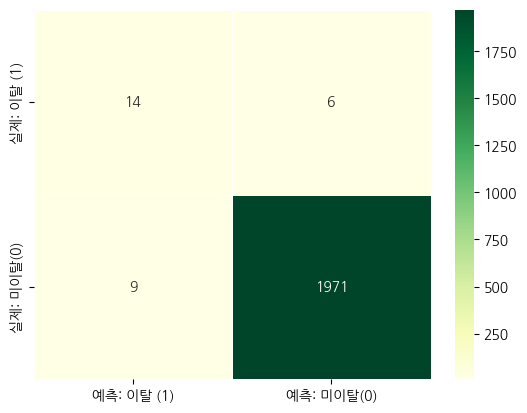
미디어 콘텐츠 미시청 행동 횟수 (총합)을 나타내는 non\_video\_viewed\_cnt\_sum 컬럼은 3000을 기준으로 총 3개의 데이터를 평균값으로 대체했다.

**모델링**

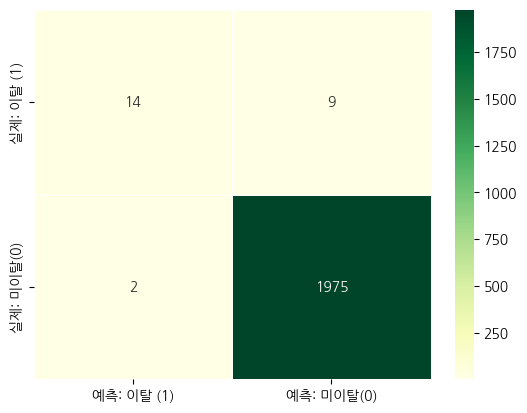
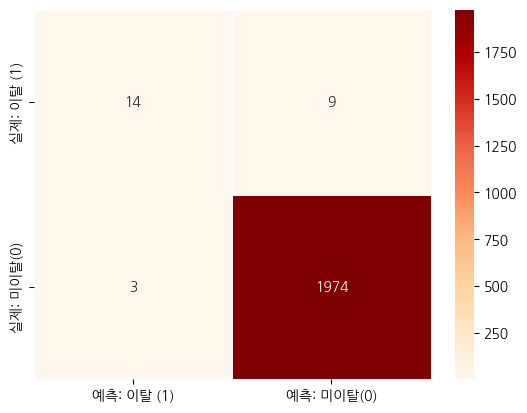
- 학습 모델은 LightGBM을 활용했으며 파라미터는 default 값으로 설정했다.

|  |  |  |
| --- | --- | --- |
| **1학년** | **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 |

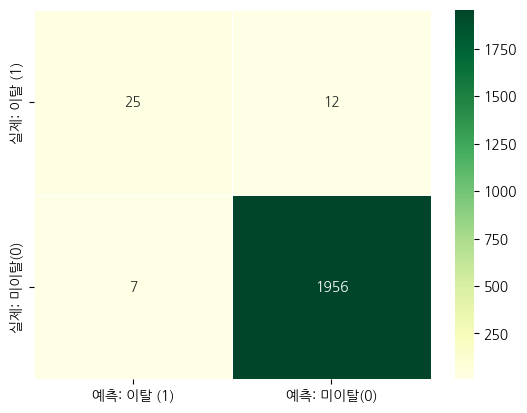
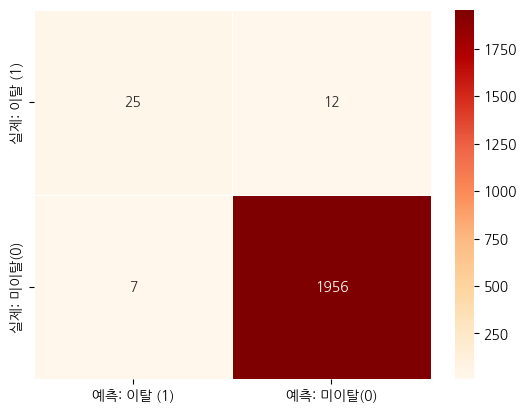




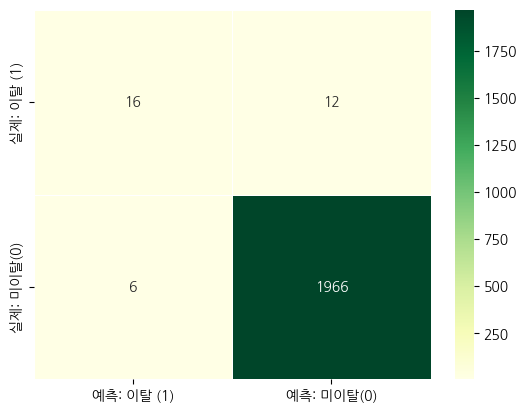
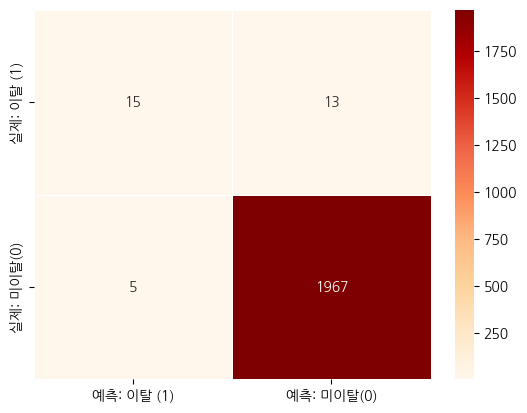
|  |  |  |
| --- | --- | --- |
| **2학년** | **Raw Data** | **Preprocess Data** |
| **Accuracy** | 0.99 | 0.99 |
| **Precision** | 0.61 | 0.61 |
| **Recall** | 0.7 | 0.7 |
| **F1-Score** | 0.65 | 0.65 |



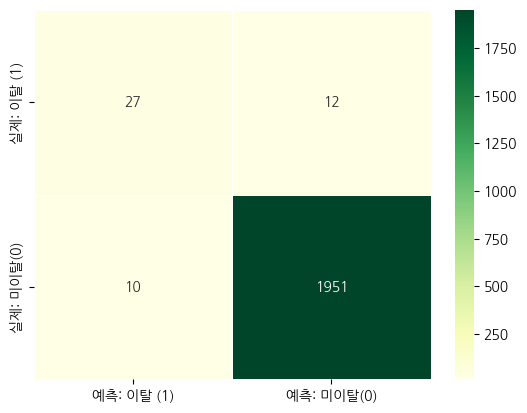
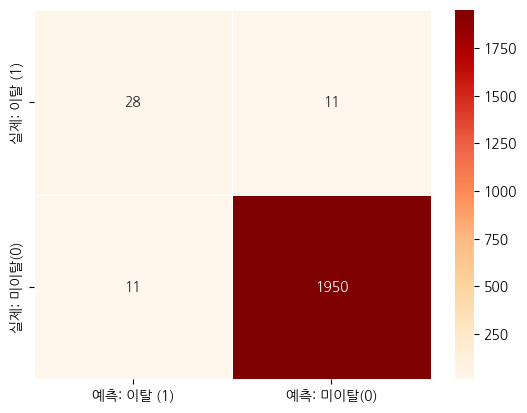
|  |  |  |
| --- | --- | --- |
| **3학년** | **Raw Data** | **Preprocess Data** |
| **Accuracy** | 0.99 | 0.99 |
| **Precision** | 0.88 | 0.82 |
| **Recall** | 0.61 | 0.61 |
| **F1-Score** | 0.72 | 0.7 |



|  |  |  |
| --- | --- | --- |
| 4학년 | **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 |



|  |  |  |
| --- | --- | --- |
| 5학년 | **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 |



|  |  |  |
| --- | --- | --- |
| 6학년 | **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 |

**결과**

**Raw Data와 Preprocess Data전반적으로 비슷한 성ㄴ**

**작업 코드**