데이터 전처리 결과서

1. 목적

- 데이터를 학습할 모델의 성능을 향상시키기 위해 학습데이터를 가공

- 결측값 보완, 피처엔지니어링, 데이터 병합 등을 사용

2. 활용 데이터

· https://www.kaggle.com/datasets/retailrocket/ecommerce-dataset

· category_tree.csv, events_csv, item_properties_part1.csv, item_properties_part2.csv

3. 데이터 전처리

· 처리 전

```
item_1 : (10999999, 4)
item_2 : (9275903, 4)
events : (2500516, 5)
category_tree : (1644, 2)
```

· 원본 데이터 컬럼

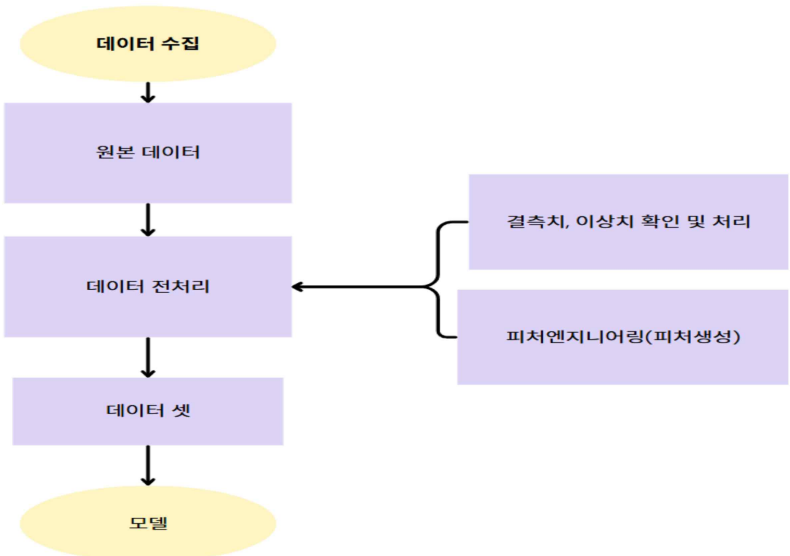
컬럼명	컬럼명(한글)	설명	데이터 타입
timestamp	타임스탬프(시각)	이벤트 발생 시각	int
visitorid	방문자ID	사용자 식별자	int
event	이벤트 종류	발생한 이벤트 유형	object
itemid	상품ID	상품 고유 식별자	int
transactionid	거래ID(구매시에만존재)	구매시의 거래의 고유 ID	float
property	속성명	이벤트와 관련된 속성	object
value	속성 값	속성의 실제 값	object
categoryid	카테고리ID	상품의 카테고리 고유 ID	int
parentid	상위카테고리ID	상품의 상위 카테고리 ID	float

· 처리 후

· 파생변수 컬럼(classification_data)

컬럼명	컬럼명(한글)	설명	데이터 타
visitorid	방문자ID	사용자 식별자 (원본 유지)	int
sessionid	방문횟수	유저의 전체 세션 수	int
item_n	구매상품수	세션별 구매 아이템 개수 평균	float
cat_n	카테고리수	세션별 고유 카테고리 수 평균	float
int_n	상호작용수	사용자가 세션에서 상호작용한 수	float
spend	총지출액	총 누적 결제 금액	float
length_min	세션길이(분)	유저의 평균 머문 시간	float
recency	마지막방문일차이(일)	마지막 방문으로부터 지난 일수	int
user_age	유저가입후경과일수	사용자	int
target_class	이탈여부	고객이 이탈(1)했는지 여부	int
ession_gap_trend	세션간격추이	세션 간 간격 표준편차	float
ctiviti_decay_ratio	최근활동감소비율	과거 활동 대비 최근 활동의 감소 비율	float
gagement_volatility	참여도변동성	세션 길이 표준편차	float
ession_interval_std	세션간격표준편차	세션 간 간격 표준편차	float
min_recency_ratio	최소리센시비율	고객 활동의 최소 리센시 비율	float
peat_category_ratio	중복카테고리클릭	동일 카테고리 반복 클릭수	float

· 전처리 흐름



· 데이터 처리 전 후

```
1 events_enh.shape
```

```
(2110041, 11)
```

```
1 classification_data.shape
```

```
(52905, 10)
```

· 처리 후 데이터 컬럼 간 상관관계