데이터 전처리 결과서

1. 수집한 데이터 설명
    a. 수집한 방법: Kaggle 데이터셋 사용
    Kaggle (Hair Salon No-Show Dataset) 출처:
    https://www.kaggle.com/datasets/frederickferguson/hair-salon-no-show-data-set/data
    b. Feature들에 대한 설명:

변수명 | 변수 설명

`#` | 예약 번호

`book_tod` | 현재 예약 시간대 (afternoon, morning, evening, Unknown)

`book_dow` | 현재 예약 요일 (Monday - Sunday)

`book_category` | 현재 예약한 서비스 (STYLE, COLOR, MISC(기타))

`book_staff` | 현재 예약 직원 (JJ, BECKY, Other)

`last_category` | 지난 예약 서비스 (Unknown, STYLE, COLOR, MISC)

`last_staff` | 지난 예약 직원 (Unknown, JJ, Other)

`last_day_services` | 지난 예약 시 받은 서비스의 수 (0 - 3)

`last_receipt_tot` | 지난 예약 때 지불한 금액

`last_dow` | 지난 예약 요일 (Monday - Sunday)

`last_tod` | 지난 예약 시간대 (afternoon, morning, evening, Unknown)

`last_noshow` | 지난 예약 때 노쇼 여부 (1: 노쇼, 0: 방문)

`last_prod_flag` | 지난 예약 시 미용실에서 상품을 구매했는가 (1: 구매, 0: 미구매)

`last_cumrev` | 고객이 지불한 누적 금액

`last_cumbook` | 고객의 누적 예약 횟수

`last_cumstyle` | 고객의 누적 예약 횟수 - STYLE

`last_cumcolor` | 고객의 누적 예약 횟수 - COLOR

`last_cumpord` | 고객의 누적 상품 구매 횟수

`last_cumcancel` | 고객의 누적 예약 취소 횟수

`last_cumnoshow` | 고객의 누적 노쇼 횟수

`noshow` | 현재 예약 노쇼 여부 (1: 노쇼, 0: 방문)

`recency` | 지난 예약 이후 고객이 방문하기까지 간격일

`first_visit` | 현재 예약이 고객의 첫 예약인지 여부 (1: 첫 예약, 0: 예약 이력 있음)

`is_revisit_30days` | 한 달 이내 재방문인지 여부. (1: 재방문, 0: 첫 방문 혹은 한 달 이후 재방문)

2. 학습 데이터에 대한 탐색적 분석 수행 결과
    a. 노쇼 비율: 전체 예약 1952건 중 노쇼는 224 건으로, 약 11%를 차지. 이는 데이터가 클래스 불균형을 가짐을 뜻하며, 학습 모델이 모두 '음성'이라고 예측해도 Accuracy는 높게 나올 수 있다고 판단. 따라서 SMOTE 기법과 같은 오버 샘플링 고려.
    b. 요일별 노쇼 경향: 일요일의 노쇼 비율이 가장 높음. 이외는 요일별로 약간의 편차는 있으나 극단적이지는 않음.
    c. 시간대별 노쇼 경향: 오후(afternoon)의 노쇼 비율이 가장 높음.
    d. 누적 노쇼 횟수와 실제 노쇼율의 관계: 누적 노쇼 횟수가 많을수록 실제 노쇼 확률도 증가.
    e. 첫 방문 여부와 예측 가능성: last_ 계열 컬럼에 결측치가 다수 존재. 이는 기록되지 않은 결측치가 아닌, 첫 방문 고객으로 판단됨. 유의미한 영향을 줄 수 있다고 판단했기에 관련된 파생 변수 생성 고려.

3. 결측치 처리 방법 및 이유
    a. 기록 누락 결측치: 최빈값으로 채움
        i. 이유: 데이터 분포의 왜곡을 최소화하고 모델이 일반적인 패턴을 학습할 수 있도록 하기 위함.
    b. last_ 계열 결측치(첫 방문임을 의미): Unknown으로 채움
        i. 유의미한 정보일 수 있는 결측값을 보존하기 위해 삭제 대신 'Unknown'으로 처리하여, 모델이 이를 하나의 특성으로 활용할 수 있도록 하기 위함.
4. 이상치 판정 기준과 처리 방법 및 이유: 요일 컬럼에서 'Monday'는 전체 데이터 중 단 1건만 존재해, 다른 요일 대비 현저히 낮은 빈도를 보였습니다. 일반적으로 범주형 변수의 경우, 특정 범주가 극단적으로 적은 비율로 나타나면 모델 학습 시 과소대표(underrepresentation)로 인해 예측 성능에 부정적인 영향을 줄 수 있습니다. 이에 따라 **'Monday'를 이상치로 간주**하였으며, 해당 데이터는 **제거**하는 방식으로 처리했습니다.
5. 적용한 Feature Engineering 방식

구분	**내용**	처리
결측치 처리	누락된 값을 대체	최빈값 또는 'Unknown'
이상치 처리	극단값 제거	특정 요일(Monday) 데이터 제거
인코딩	범주형 → 수치형 변환	Label Encoding, One-Hot Encoding
스케일링	수치형 정규화	StandardScaler
파생 변수 생성	새로운 컬럼 만들기	first_visit: 첫 방문인지 여부 is_revisit_30days: 한 달 내 재방문인지 여부
오버 샘플링	SMOTE 기법	소수 클래스(노쇼) 증폭