데이터 전처리 결과서

목차

---

1. 수집한 데이터 설명

Hair Salon No-Show Dataset

- 출처: Kaggle (https://www.kaggle.com/datasets/frederickferguson/hair-salon-no-show-data-set/data)
- 관측치 수: 1952
- 컬럼 수: 22

프로젝트에서 사용한 데이터셋은 Hair Salon No-show Dataset으로, 캐글에서 가져왔다.
본 데이터는 캐나다 토론토에 위치한 실제 미용실의 허가를 받아 사용되었다. 이 데이터는 “어떤 고객 예약이 노쇼 가능성이 가장 높은가”라는 질문에 답하기 위해 수집되었으며, 2018년 3월부터 7월까지의 시계열 데이터를 기반으로 한다.

X 데이터는 기존 두 개 변수를 제거하고(예약 번호(#)와 노쇼 여부(noshow)), 새로운 파생 변수 두 개를 추가하여(첫 방문 여부(first_visit)와 30일 이내 재방문 여부 (is_revisit_30days)) 총 22개의 열로 구성되었다. target 변수는 노쇼 여부(noshow)이다.

변수 설명

변수명	변수 설명	변수 자료형	구성된 값

#	예약 번호	범주형	0 ~ 1951
book_tod	현재 예약 시간대	범주형	afternoon, morning, evening, Unknown
book_dow	현재 예약 요일	범주형	Monday - Sunday
book_category	현재 예약한 서비스	범주형	STYLE, COLOR, MISC(기타)
book_staff	현재 예약 직원	범주형	JJ, BECKY, Other
last_category	지난 예약 서비스	범주형	Unknown, STYLE, COLOR, MISC
last_staff	지난 예약 직원	범주형	Unknown, JJ, Other
last_day_services	지난 예약 시 받은 서비스의 수	수치형	0 ~ 3
last_receipt_tot	지난 예약 때 지불한 금액	수치형	0 ~ 383
last_dow	지난 예약 요일	범주형	Monday - Sunday
last_tod	지난 예약 시간대	범주형	afternoon, morning, evening, Unknown
last_noshow	지난 예약 때 노쇼 여부	이진형	1: 노쇼, 0: 방문
last_prod_flag	지난 예약 시 상품 구매 여부	이진형	1: 구매, 0: 미구매
last_cumrev	고객이 지불한 누적 금액	수치형	0 ~ 1276
last_cumbook	고객의 누적 예약 횟수	수치형	0 ~ 20
last_cumstyle	고객의 누적 예약 횟수 - STYLE	수치형	0 ~ 18
last_cumcolor	고객의 누적 예약 횟수 - COLOR	수치형	0 ~ 7
last_cumpord	고객의 누적 상품 구매 횟수	수치형	0 ~ 11
last_cumcancel	고객의 누적 예약 취소 횟수	수치형	0 ~ 8
last_cumnoshow	고객의 누적 노쇼 횟수	수치형	0 ~ 9
noshow	현재 예약 노쇼 여부	이진형	1: 노쇼, 0: 방문
recency	지난 예약 이후 방문 간격일	수치형	0 ~ 133

`first_visit` (파생 변수)	첫 예약 여부	이진형	1: 첫 예약, 0: 예약 이력 있음
`is_revisit_30days` (파생 변수)	한 달 이내 재방문 여부	이진형	1: 재방문, 0: 첫 방문/한 달 이후 재방문

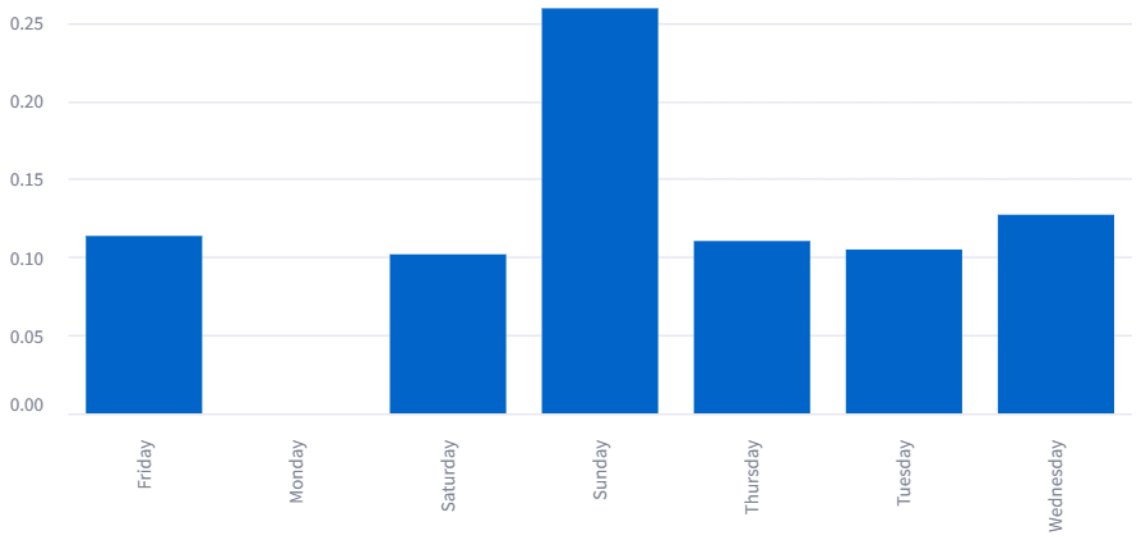
---

2. 학습 데이터에 대한 탐색적 데이터 분석(EDA) 수행 결과

a. 노쇼 비율

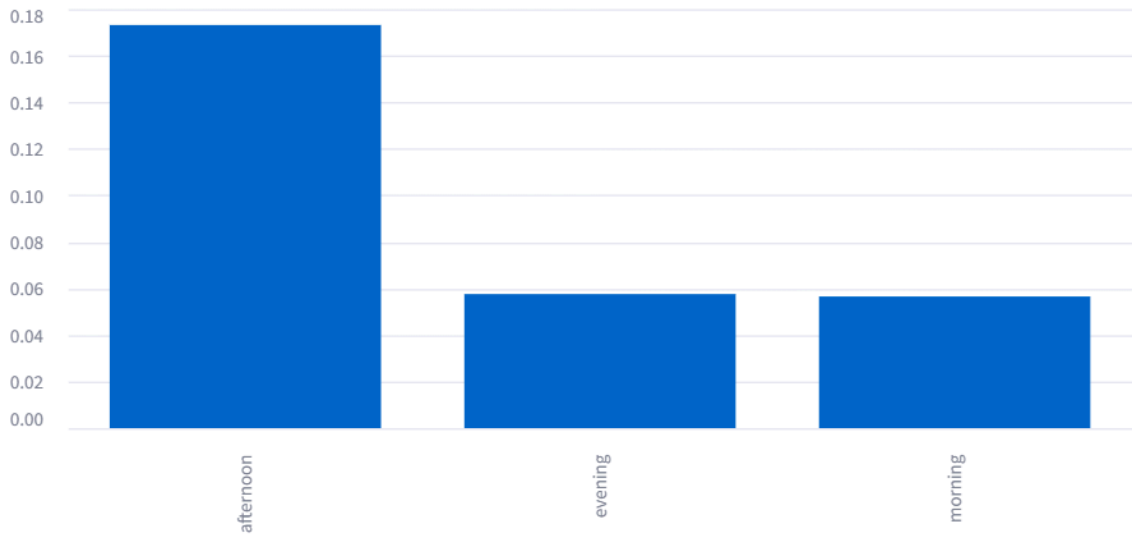
전체 예약 1952건 중 '노쇼'(양성 클래스)는 224건으로, 약 11%를 차지한다. 이는 데이 터가 클래스 불균형임을 뜻하며, 예측 모델이 모든 데이터를 '음성'(노쇼를 하지 않음을 뜻함)이라고 예측해도 정확도(accuracy)는 높게 나올 것이라고 판단하였다. 따라서 데이터 불균형 문제를 해소를 위해 SMOTE와 같은 오버 샘플링을 고려하였다.

b. 요일별 노쇼 경향



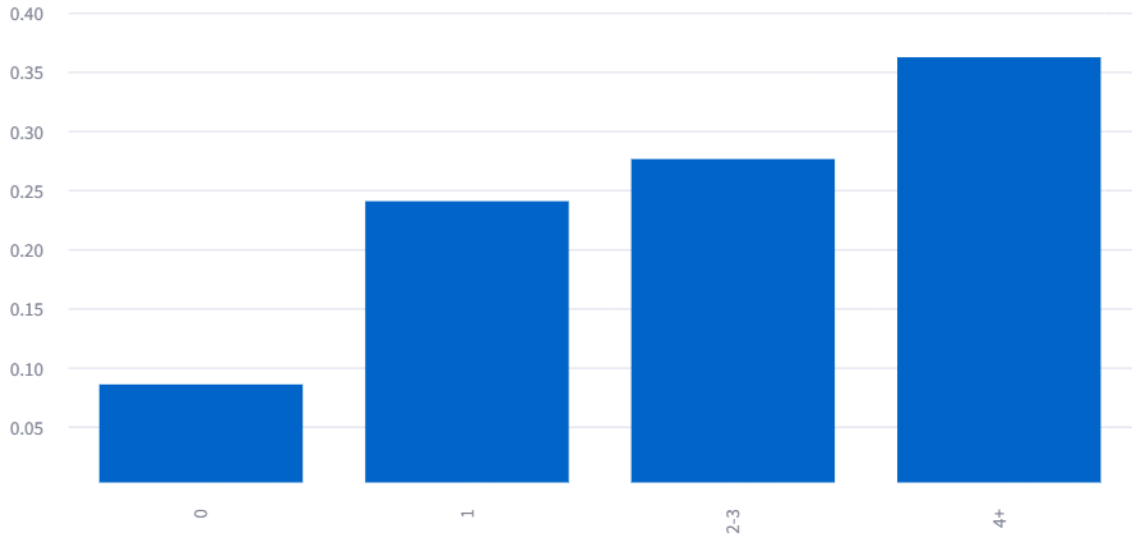
- 일요일에 노쇼 비율이 가장 높다.
- 일, 월을 제외한 나머지 요일의 경우 대체로 비슷하다.

c. 시간대별 노쇼 경향



- 오후(afternoon)에 노쇼 비율이 가장 높다.
- 저녁(evening)과 아침(morning) 시간의 경우 큰 차이가 없다.

d. 누적 노쇼 횟수와 현재 예약 노쇼 비율의 관계



- 누적 노쇼 횟수를 x축, 현재 예약 노쇼 비율을 y축으로 하고 시각화한 결과, 누적 노쇼 횟수가 많을수록 실제 노쇼할 가능성이 높아진다.
  - 누적 0회 → 노쇼율 매우 낮음
  - 누적 2~3회 → 점진적 증가
  - **누적 4회 이상 고객**의 노쇼율은 평균 대비 **3배 이상**

e. 파생 변수 생성

- 첫 방문 여부 - `first_visit`
  - `last_` 계열 컬럼에 결측치가 다수 존재하였다. 이는 기록되지 않은 결측치가 아닌 '첫 방문 고객'으로 판단된다. 첫 방문 여부가 노쇼 여부 판단에 유의미한 단서가 될 수 있다고 판단해 파생 변수를 생성했다.
- 30일 이내 재방문 여부 - `is_revisit_30days`
  - 지난 예약 이후 현재 예약까지의 방문 간격을 나타내는 변수 `recency` 를 기반으로, 30일 이내 재방문 여부를 나타내는 파생 변수를 생성했다.

---

3. 결측치 처리

- 기록 누락 결측치는 최빈값으로 채워 데이터 분포의 왜곡을 최소화하고 모델이 일반적인 패턴을 학습할 수 있게 했다.
- 지난 방문에 대한 정보를 나타내는 `last` 계열 변수가 결측치라는 것은 해당 예약이 고객의 첫 예약임을 의미한다. 결측치라는 것 자체가 유의미한 의미를 가지므로, 행을 삭제하는 대신 결측치를 'Unknown'으로 채웠다.

---

4. 이상치 처리

요일 컬럼에서 'Monday'는 전체 데이터 중 단 한 건만 존재해, 다른 요일 대비 현저히 낮은 빈도를 보였다. 일반적으로 범주형 변수의 경우, 특정 범주가 극단적으로 적은 비율로 나타나면 모델 학습 시 과소대표(Underrepresentation)로 인해 예측 성능에 부정적인 영향을 줄 수 있다. 따라서 'Monday' 를 이상치로 간주하고 해당 데이터를 제거했다.

---

5. 적용한 Feature Engineering 방식

구분	내용	처리
결측치 처리	누락된 값을 대체	최빈값 또는 'Unknown'
이상치 처리	극단값 제거	특정 요일(Monday) 제거
인코딩	범주형 변수를 모델이 학습하기 용이한 형태로 변환	Label Encoding, One-Hot Encoding
스케일링	수치형 변수의 값 범위 조정	StandardScaler
파생 변수 생성	새로운 컬럼 생성	`first_visit` : 첫 방문인지 여부 `is_revisit_30days` : 한 달 내 재방문인지 여부
오버 샘플링	SMOTE	소수 클래스(노쇼) 증폭