24.11.30 (Hypothesis_testing)

Hypothesis_testing 진행 계획 및 variable type분석

- 60210021 박세민

1. Hypothesis_testing 목표: 가설 검정을 통해 회귀분석할 freature 선정하기

아래의 두 가지 케이스에 대해 각각 가설검정을 수행할 예정

1. 종속변수가 범주형일 때 (logistic regression을 위한 feature 선정)

logistic regression 가설: 해당 피처들을 기반으로 어떤 종류의 호텔(hotel: city hotel/resort hotel) 선호할 지 예측 -> city hotel/resort hotel을 기준으로 연속형 컬럼들과 two sample t-test/카이제곱 test 수행(독립성/정규성/등분산성 확인 필요)

2. 종속변수가 연속형일 때 (linear regression을 위한 feature 선정)

linear regression 가설: 해당 피처들을 기반으로 호텔을 예약할 경우 예상 일일 숙박 비용 예측 -> adr(평균 일일 숙박비용)을 기준으로 연속형 컬럼들과 피어슨 상관관계 분석 진행

2. variable type 판단하기

데이터셋 컬럼 정보(총 32개)
hotel: 호텔의 종류(Resort Hotel / City Hotel) - Nominal
is_canceled: 호텔 예약 취소 여부(1: 취소함 / 2: 취소 안함) - Binary(Nominal)
lead_time: 호텔 예약 날짜와 호텔 도착 날짜 사이 경과 일수 - Numeric
arrival_date_year: 호텔에 도착한 년도 – Date(Ordinal)
arrival_date_month: 호텔에 도착한 월 – Date
arrival_date_week_number: 호텔에 도착한 주 - Date
arrival_date_day_of_month: 호텔에 도착한 일 – Date
stays_in_weekend_nights: 주말에 숙박한 일 수 – Numeric(범주형 변환 가능)
stays_in_week_nights: 주중에 숙박한 일 수 - Numeric(범주형 변환 가능)
adults: 호텔에 방문한 어른 수 - Numeric
children: 호텔에 방문한 청소년 수 - Numeric
babies: 호텔에 방문한 유아 수 - Numeric
meal: 예약된 식사 유형(Undefined / SC는 식사가 없는 경우) - Nominal
country: 호텔을 예약한 손님의 국가 - Nominal
market_segment: 마켓 구분(TA: Travel Agents / TO: Tour Operator) (아래 컬럼이랑 비슷한 느낌인거 같은데, 더 찾아보겠습니다) - Nominal
distrubution_channel: 예약 채널(TA: Travel Agents / TO: Tour Operator) (생산자에서 소비자에게 이르기까지의 방식, 채널) - Nominal
is_repeated_guest: 호텔 재방문 손님 여부(1: 맞음 / 2: 아님) - Binary(Nominal)
previous_cancellations: 해당 손님의 과거 예약 취소 횟수 - Numeric
prevoius_bookings_not_canceled: 해당 손님의 과거 예약 취소하지 않은 횟수 - Numeric
reserved_room_type: 예약된 객실 유형(익명성을 위해 명칭 대신 코드로 표시 'C' 'A' 'D' 'E' 'G' 'F' 'H' 'L' 'P' 'B') - Nominal
assigned_room_type: 예약에 배정된 객실 유형('C' 'A' 'D' 'E' 'G' 'F' 'I' 'B' 'H' 'P' 'L' 'K') - Nominal
booking_changes: 예약 후 예약 변경 혹은 수정 횟수 - Numeric
deposit_type: 보증금 지불 여부('No Deposit', 'Refundable', 'Non Refund') - Nominal
agent: 호텔 예약 여행 에이전시 ID - ID
company: 호텔 예약 담당 회사 ID - ID
days_in_waiting_list: 예약이 완료될 때까지 대기 목록에 속한 기간 - Numeric
customer_type: 고객 유형('Transient', 'Contract', 'Transient-Party', 'Group') - Nominal
adr: 평균 일일 숙박비용 - Numeric
required_car_parking_spaces: 고객이 요구한 주차공간 수 - Numeric
total_of_special_requests: 고객이 요청한 특별 요청 수(예: 트윈 침대, 높은 층수) - Numeric
reservation_status: 마지막 예약 상태('Check-Out', 'Canceled', 'No-Show') - Nominal
reservation_status_date: 마지막 예약 상태가 설정된 날짜 – date

3. 독립변수를 선택하는 기준

1. 도메인 지식 기반 선택
 : 문제의 배경과 목표: 분석의 목적과 관련이 깊은 변수를 우선적으로 선택합니다.
   예: 학생 성적 예측 모델에서는 공부 시간, 출석률 등이 중요할 가능성이 높습니다.
2. 독립변수와 종속변수 간의 상관관계 분석(상관계수나 t-test, anova분석, 카이제곱 분석)
3. 상관관계가 높은 변수 제거
 : 독립변수 간 상관관계가 높은 경우, 대표 변수를 선택
4. (regression 파트=>) 회귀분석에 적합한 변수 스케일링, 로그 변환