**문제 유형6**

여행자 보험 청구 여부를 예측하는 모델을 개발하고자 한다. 데이터셋은 다음과 같다.

travel\_insurance2.csv (구분자: 쉼표(,), 63,326 Rows, 12 Columns, UTF-8 Encoding, 결측치 존재(Null, N/A, Empty Cell)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **변수명** | **설명** | **타입** | **변수명** | **설명** | **타입** |
| NO | ID | Double | Duration | 구매자 여행 기간 | Double |
| Agency | 여행상품 판매 기관 | String | Destination | 구매자 여행 국가 | String |
| AgencyType | 여행상품 판매처 타입 | String | NetSales | 순매출액 | Double |
| DistributionChannel | 유통 채널 | String | Commision | 수수료율(%) | Double |
| ProductName | 상품명 | String | Gender | 구매자 성별 | String |
| Claim | 보험 청구 여부(Yes/No) | String | Age | 구매자 연령 | Double |

**\* 분석 수행하기 전, 다음 전처리 단계들을 순서대로 수행하시오.**

**단계 1-1** ProductName, NetSales 변수를 제거하시오.

**단계 1-2**. 성별(Gender)에 결측치가 존재한다. 다음 규칙에 따라 성별의 결측치를 보정하시오.

If (NO=짝수) then (Gender=’M’)

Else if (NO=홀수) then (Gender=’F’)

Else 성별(Gender)이 결측치가 아닌 경우 현재 값 유지

**\* 상기 전처리 완료 후(데이터 프레임명: basetable1) 다음 분석(문제 1~3)을 수행하시오.**

1. (basetable1 이용) 40대(40<=Age<50)이고 여성(Gender=’F’)인 여행자가 두 번째로 많이 방문하는 여행지 국가(Destination)를 기술하시오(여행지 국가(Destination)는 대문자로 기술, 예시: ITALY)

|  |  |
| --- | --- |
| 정답 | MALAYSIA |

2. (basetable1 이용) 유통 채널(DistributionChannel)에 따라 여행 기간(Duration)의 평균이 통계적으로 차이가 있는지 적절한 검정을 수행한 후, 검정 통계량(t-value)의 절댓값을 기술하시오.

- 등분산 가정하고 유의수준 0.05 하에서 양측검정

- 검정 통계량(t-value)의 절댓값은 소수점 둘째 자리까지만 기술(이후 자리는 버림, 예시: 0.12)

|  |  |
| --- | --- |
| 정답 | 10.289600869594274 |

3. (basetable1 이용) 마지막으로 청구(Claim) 여부를 예측하는 로지스틱 회귀 모형을 만들고자 합니다. 다음의 전처리를 수행한 후 분석을 수행하시오.

**단계 1-3**. 범주형 변수인 성별(Gender)와 유통 채널(DistributionChannel)에 대한 가변수(총 2개)를 추가하시오.

- 가변수화 Hint

(Brightics Studio) : One Hot Encoder 사용, Drop Last=True

(Python) : pd.get\_dummies() 함수 사용, drop\_first=True

**단계 1-4**. 다음과 같이 Train과 Test 데이터셋을 분리하시오.

- Train DataSet: NO가 3의 배수가 아닌 데이터

- Test DataSet: NO가 3의 배수인 데이터

**단계 1-5**. Train DataSet으로 로지스틱 회귀분석 모델(Logistic Regression Model)을 학습하시오.

- 독립변수(총 5개): Duration, Commision, Age,

성별(Gender)과 유통 채널(DistributionChannel)의 가변수(단계 1-3.에서 생성된 변수)

- 종속변수: 청구 여부(Claim)

- 함수 가이드(Brightics Studio)

|  |
| --- |
| Inverse of Regularization=100,000, Seed=1234, Penalty: l2, Solver: newton-cg, 나머지: Default |

**단계 1-6**. **단계 1-5.**에서 학습한 모델을 **단계 1-4.**에서 분리한 ‘Test DataSet’에 적용하여 청구 확률 (Claim=Yes 일 확률)을 구한 후, Lift 변수를 추가하시오.

Lift = 청구 확률/0.015

**(※ 0.015: 전체 데이터 셋에서 Claim=Yes인 비율, 반올림 적용)**

데이터셋을 청구 확률 (Claim=Yes 일 확률) 역순으로 (내림 차순, DESC ORDER) 정렬한 뒤, 상위 100등 안쪽(<=, 이하)에 포함되는 데이터들에 대한 Lift의 합계를 기술하시오.(Lift의 합계는 정수 부분만 기술(예시: 12)

- Hint(Brightics Studio) : rank() over (partition by A order by B desc/asc) 활용

|  |  |
| --- | --- |
| 정답 | 1344.105337707629 |