**문제 유형2**

보험사에서 고객(총 29,132명) 신상 정보와 21개의 보험 상품 구매 정보를 이용하여 보험 상품 추천 시스템을 개발하고자 하며, 제공된 데이터셋 정보는 다음과 같다.

Train2.csv (구분자: “,”(Comma), 29,132 Rows, 26 Columns, UTF-8 Encoding)

|  |  |  |
| --- | --- | --- |
| **변수명** | **설명** | **데이터 타입** |
| ID | 고객 ID | String |
| join\_date | 가입 일자, DAY/MONTH/YEAR 형식  예: 2021년 4월 8일 가입->8/4/2021 | String |
| sex | 성별(M/F) | String |
| birth\_year | 출생 연도 | Double |
| branch\_code | 모집 사무소 | String |
| P1~ P21 | 보험 상품 1~21에의 가입 여부(1: 가입, 0: 미가입) | Double |

**\* 다음의 전처리를 먼저 수행한 후 분석 진행하시오.**

**단계 1-1**. 가입 일자(join\_date) 컬럼의 연도 정보를 이용하여 가입 연령(변수명: reg\_age) 컬럼을 추가하시오(join\_date)를 기준으로 결측치를 제거한 후 변수 생성하시오).

reg\_age = 가입 연도 - 출생 연도 + 1

**단계 1-2**. 가입한 보험 상품의 수가 4개 이상(>=)인 고객은 VIP로 분류하는 VIP 컬럼(변수명: VIP)을 추가하시오.

IF 가입한 보험 상품의 수 >= 4 then ‘VIP’ else ‘Not VIP’ (VIP는 총 1,231명)

(가입한 보험 상품 수는 P1~P21 상품 중 가입한 상품의 수를 집계한 후 사용)

**\* 상기 전처리를 완료한 데이터 프레임(데이터 프레임명: basetable1, 29132(결측치 포함)/29130(결측치 제거) Rows)으로 다음 분석(문제 1~3) 수행**

1. (basetable1을 이용) 모집 사무소(branch\_code) 별 가입 연령의 평균을 집계하고, 이 값이 가장 큰 모집 사무소를 기술하시오. (답안 예시) ABC

|  |  |
| --- | --- |
| 정답 | 49BM |

2. (basetable1을 이용) 고객의 VIP 여부가 성별(sex)과 서로 독립이라 할 수 있는지 적절한 검정을 수행하고, 검정 결과 p-value를 기술하시오.

- p-value는 소수점 셋째 자리 이하는 버리고 소수점 둘째 자리까지 기술 (답안 예시) 0.1234

|  |  |
| --- | --- |
| 정답 | 0.04254096311321683 (결측치 포함)  0.0442845227544766 (결측치 제거) |

3. (basetable1을 이용) 고객들이 어떤 보험 상품들을 같이 가입했는지 연관성 분석을 다음 조건에 따라 수행하고자 한다.

- 21 개의 보험 상품(P1~P21 활용)을 분석 대상으로 하며, 1개의 보험상품에만 가입한 고객은 분석에서 제외

- Hint

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| cust\_id | P1 | P2 | P3 | → | (가입한 상품) |
| 4WK | 1 | 1 | 0 | {“P1”, “P2”} |
| CP5 | 0 | 1 | 1 | {“P2”, “P3”} |

- (Association Rule 모델 가이드) Min Support: 0.01, Min Confidence: 0.01, 그 외: Default

연관성 분석 결과 중, 선행(antecedent)과 후행(consequent)이 단일 아이템으로 구성된 rule만 추출하시오. 이 결과를 활용하여, 보험상품 'P15'에 가입한 고객에게 추천하기 가장 적합한 보험 상품을 Lift 기준(내림차순)으로 선정하고, 그 때의 Lift를 기술하시오.

- Lift는 소수점 넷째 자리 이하는 버리고 셋째 자리까지만 기술하시오.

(답안예시) 1.234

|  |  |
| --- | --- |
| 정답 | 1.30791769238008 (결측치 포함)  1.3077626374413094 (결측치 제거) |