**문제 유형4**

실시간 중계 Tweet이 방문자들의 반응지표들과 관계가 있는지에 대한 분석을 수행하고자 한다. 제공된 데이터셋 정보는 다음과 같다.(2017년 9월 29일부터 10 월 26일까지 개인 계정에서 보낸 313개의 트윗 정보 수집)

**tweets2.csv** (구분자: “,”(comma), 313 Rows, 13 Columns, UTF-8 Encoding):

|  |  |  |
| --- | --- | --- |
| **변수명** | **설명** | **데이터 타입** |
| No | Index | Double |
| atConference | 실시간 중계 여부(TRUE/FLASE) | String |
| day | 트윗 일자(YYYY-MM-DD) | String |
| impressions | 트윗 조회 수 | Double |
| retweets | 리-트윗 횟수 | Double |
| likes | “좋아요” 클릭수 | Double |
| userprofileclicks | 프로필 조회 횟수 | Double |
| urlClicks | URL 클릭 횟수 | Double |
| hashtagClicks | 트윗에서 해시 태그를 클릭한 횟수 | Double |
| detailExpands | 트윗 확장 횟수 | Double |
| follows | Follower 수 | Double |
| mediaViews | 트윗에 포함된 미디어를 본 횟수 | Double |
| mediaEngagements | 트윗에 포함된 미디어를 클릭한 횟수 | Double |

**\* 다음의 전처리를 먼저 수행한 후 진행하시오.**

**단계 1-1**. atConference, day, impressions를 제외한 나머지 9개 변수의 합 변수(변수명: totalSum)를 추가하시오.

**\* 상기 전처리를 완료한 데이터 프레임(데이터 프레임명: basetable1)으로 다음 분석(문제 1~3)을 수행하시오.**

1. (basetable1을 이용) 리-트윗 횟수(retweets)가 0이 아닌 트윗을 대상으로 실시간 중계 여부(atConference)에 따라 좋아요(‘likes’) 클릭 수에 평균 차이가 있는지 적절한 검정을 수행하고 검정 결과의 t-statistic를 기술하시오.

- 주어진 데이터가 정규성과 등분산 조건을 만족한다는 가정을 하며, t-statistic은 절대치를 취한 후 소수점 다섯째 자리 이하는 버리고 넷째 자리까지만 기술

|  |  |
| --- | --- |
| 정답 | 4.702598503830561 |

\* 실시간 중계 여부(atConference)를 예측하는 모델을 생성하기 위해 **basetable1**에 다음의 전처리를 수행하시오.

**단계 1-2**. No, atConference, day, totalSum 를 제외한 10개 변수에 Min-Max Scaler를 적용하시오.

**단계 1-3**. day변수를 이용하여 요일변수(변수명: weekday, 월요일/화요일/…/일요일)를 추가하시오.

* 파이썬의 경우 Series.dt.day\_name(locale=’ko\_kr’) 함수 활용

**단계 1-4.** 다음 조건에 따라 train과 test데이터셋을 생성하시오.

* train: No가 4의 배수가 아닌 데이터, test: No가 4의 배수인 데이터

2. 실시간 중계 여부(atConference)를 예측하는 K-NN 분류 모델을 만들고자 한다.

* 독립 변수(총 10개): Min-Max Scaler를 적용한 변수 사용
* 분류 기준: Euclidean 거리를 이용하여 가까운 5개 이웃(Neighbor)의 실시간 중계 여부(atConference) 참조

도출된 KNN 분류 모델을 통하여 Test 데이터셋의 실시간 중계 여부(atConference) 예측 시, 실제로는 실시간 중계지만(atConference=”TRUE”) 실시간 중계가 아닌 것으로 예측한(predicted=”FALSE”) 데이터의 수는 몇 개인가?

|  |  |
| --- | --- |
| 정답 | 2 |

3. (단계1-4에서 생성된 train/test dataset 활용) 실시간 중계 여부(atConference)를 예측하는 로지스틱 회귀 분류 모델을 만들고자 한다.

* 독립 변수(16개) : No, atConference, day, totalSum를 제외한 수치형 변수(minmax 표준화 적용), weekday의 더미 변수(6개)
* 수치형 변수: 'impressions', 'retweets', 'likes', 'userprofileclicks', 'urlClicks', 'hashtagClicks', 'detailExpands', 'follows', 'mediaViews', 'mediaEngagements'
* 모델 가이드 : Inverse of Regularization= 100000, seed=123, Penalty: l2, solver: newton-cg
* 모델 도출 시 train 데이터셋 사용

도출된 모델에 test 데이터셋을 적용하여 실시간 중계 여부(atConference)가 “TRUE”인 데이터에 대한 Log Probability TRUE의 합계를 구하시오.(절대값을 취한 후 소수점 다섯째 자리 이하는 버리고 넷째 자리까지만 기술)

* Brightics와 파이썬 패키지 버전 차이로 약간의 차이가 발생할 수 있음

|  |  |
| --- | --- |
| 정답 | 2.616505654313135, (2.88) |