**문제 유형5**

* 휴대폰 음질 측정 데이터로 음질의 특징을 분석하고자 한다. MOS(Mean Opinion Score) 변수는 두 지점간의 통신망과 테스트 장비를 이용하여 측정한 음질 측정 지표이며, 다음은 데이터셋에 대한 정보이다.

**phone.csv** (구분자: “,” , 105,828 Rows, 10 Columns)

|  |  |  |
| --- | --- | --- |
| **변수명** | **설명** | **데이터 타입** |
| Date | 테스트 실행 시점 | String |
| Signal | 테스트 시 측정된 신호 강도 (단위: dBm) | double |
| Speed | 테스트 시 측정된 전송 속도 (단위: m/s) | double |
| Distance | 테스트 시 두 지점 간의 거리 (단위: m) | double |
| Manufacturer | 제조사 | String |
| TestResult | 테스트 결과 (SUCCESS/FAILURE - DROP CALL/FAILURE - SETUP FAIL) | String |
| TestTechnology | 테스트 시 사용된 이동 통신망 유형 (GSM/LTE/UMTS) | String |
| SetupTime | 테스트 시 소요된 통화 연결 시간(s) | double |
| MOS | 테스트 시 측정된 음질 측정 지표 | double |
| Acceptability | 테스트 시 서비스 만족여부(0[불만족/1[만족]/nan[결측치]] | double/string |

**\* 다음의 전처리를 먼저 수행한 후 분석 진행하시오.**

**단계 1-1**. 데이터 중 테스트 결과(TestResult)가 SUCCESS이고 신호 강도(Signal) 값이 결측치(null, N/A 등)가 아닌 데이터만 추출

**단계 1-2**. 결측치 보정

1. 두 지점 간의 거리(Distance) 변수의 결측치는 통신망 유형별(TestTechnology)로 두 지점 간의 거리(Distance) 값이 결측치가 없는 데이터로 다음과 같은 선형식을 만들고 이를 이용하여 결측치를 보정(OLS(Ordinary Least Squares) 적용)

* 모델 생성 시 거리(Distance) 변수를 기준으로 결측치가 없는 데이터를 train용으로 구성
* 결측치 보정 시 거리(Distance) 변수를 기준으로 결측치가 있는 데이터를 생성된 모델에 적용해서 보정
* 이때 생성된 모델은 통신망 유형(TestTechnology)별로 생성되어야 하며, 결측치 보정 시에도 적용해야 함

1. 통화 연결 시간(SetupTime)이 0인 데이터는 통신망 유형(TestTechnology)별로 ‘통화 연결 시간이 0이 아닌 데이터들의 평균 값’으로 보정
2. 전송 속도(Speed)가 0 이하인 데이터는 1) 단계와 동일하게 통신망 유형별(TestTechnology)로 전송 속도(Speed) 값이 0 초과(>)인 데이터를 사용하여 선형식을 만들고 이를 이용하여 추정해 보정

* 모델 생성 시 전송 속도(Speed) 변수를 기준으로 0초과인 데이터를 train용으로 구성
* 전송 속도(Speed) 0 이하인 데이터 보정은 전송 속도(Speed) 변수 기준으로 0이하인 데이터를 생성된 모델에 적용해서 보정
* 이때 생성된 모델은 통신망 유형(TestTechnology)별로 생성되어야 하며, 전송 속도(Speed) 0 이하인 데이터 보정 시에도 반영해서 보정해야 함

**\* 상기 전처리를 완료한 데이터 프레임(데이터 프레임명: basetable1, 105,147 rows)으로 다음 분석(문제 1~3) 진행하시오.**

1. (basetable1 이용) 통신망 유형(TestTechnology)별로 통화 연결 시간의 이상치(Outlier)를 제거한 후 통신망 유형(TestTechnology)별로 통화 연결 시간(SetupTime)에 차이가 있는지 검정하고 검정 통계량(F Statistic)을 기술하시오.(이때 이상치는 Tukey 방식 적용하며, 검정 통계량은 정수부분만 기술)

- 이때 집단 간의 평균 차이 검정 시 정규성과 등분산을 가정함

- [Tukey 방식]

Q1: 제 1사분위수, Q3: 제 3사분위수, IQR: Q3 – Q1 일 때,

(Y < Q1 - 1.5\*IQR) or (Y > Q3 + 1.5\*IQR) 이면 Y 이상치

|  |  |
| --- | --- |
| 정답 | 891661.2777499713 |

**단계 1-3**. (basetable1 이용) 테스트 시점의 날짜(Date)의 요일 변수(Day) 생성 후 요일 변수(Day)와 통신망 유형(TestTechnology) 변수의 가변수(Dummy)를 생성하시오.  
 - Brightics의 경우 drop\_last=True, 파이썬의 경우 drop\_first=True 적용

**단계 1-4**. (단계1-3. 이용) Split Data 함수를 사용하여 Train과 Test 데이터를 분할하시오.

* Train:Test = 7:3 비율, 음질의 만족 여부(Acceptability)에 따라 층화추출 적용, seed(random\_state) : 123
* Python인 경우 Acceptability 변수를 기준으로 결측치를 제외한 데이터셋으로 train과 test 분할하며, 이때 sklearn.model\_selection.train\_test\_split() 함수 사용

2. (단계 1-3. 생성된 데이터셋 이용) 음질의 만족 여부(Acceptability)를 예측하는 KNN 분류 모델을 만드시오.

- Train Data 사용

- 독립변수(총 6개): 신호 강도(Signal), 전송 속도(Speed), 두 지점 간의 거리(Distance), 통화 연결 시간(SetupTime), 통신망 유형(TestTechnology)의 가변수(2개)

생성된 모델에 Test 데이터를 적용하여 음질의 만족 여부(Acceptability)를 예측한 후 모델의 성능평가로 AUC(the Area Under a ROC Curve) 값을 기술하시오.

- KNN 분류 모델 가이드 : K=5

- AUC 값은 소수점 셋째 자리 이하는 버리고 둘째 자리까지 기술

|  |  |
| --- | --- |
| 정답 | 0.93 (0.91) |

3. 음질의 만족 여부(Acceptability)를 예측하는 모델을 Decision Tree 알고리즘을 이용하여 만들고자 한다. 통신망 유형(TestTechnology)별로 분리해서 모델을 다음 조건에 따라 만드시오.

- Train Data 사용

- 독립변수(총 10개): 신호 강도(Signal), 전송 속도(Speed), 두 지점 간의 거리(Distance), 통화 연결 시간(SetupTime), 요일의 가변수(6개)

이 모델에 Test데이터를 적용하여 음질의 만족 여부(Target)를 예측한 후 정확도(Accuracy) 값을 기술하시오.

- [Decision Tree 모델 가이드]

분순도 기준: Gini, Max Depth: 6, Min Samples Splits: 5, Seed: 1234, 그 외: Default

- 정확도(Accuracy) 값은 소수점 둘째 자리까지 기술, 이후 자리는 절삭.

|  |  |
| --- | --- |
| 정답 | 0.90 |