**DS 실기 공개 문항**

**2021.03**

**ProDS 인증위원회**

**□ 대문항 세트1**

마케팅 전략을 수립하기 위해 신용 카드 고객을 대상으로 고객 세분화(Customer Segmentation) 및 예측 모델링을 수행하고자 한다.

**DS\_Sample\_1.csv** (구분자: comma(“,”), 1,000 Rows, 18 Columns, UTF-8 인코딩)

|  |  |  |
| --- | --- | --- |
| **컬럼** | **정의** | **Type** |
| CUST\_ID | 고객 ID | Double |
| BALANCE | 연간 평균 잔고액 | Double |
| BALANCE\_FREQUENCY | 연중 잔고액 갱신 개월 수 비율 (0~1 사이값) | Double |
| PURCHASES | 구매 총액 | Double |
| ONEOFF\_PURCHASES | 일시불 구매 총액 | Double |
| INSTALLMENTS\_PURCHASES | 할부 구매 총액 | Double |
| CASH\_ADVANCE | 현금서비스 구매 총액 | Double |
| PURCHASES\_FREQUENCY | 연중 구매 개월 수 비율 (0~1 사이값) | Double |
| ONEOFF\_PURCHASES\_FREUQUENCY | 연중 일시불 구매 개월 수 비율 (0~1 사이값) | Double |
| PURCHASES\_INSTALLMENTS\_FREQUENCY | 연중 할부 구매 개월 수 비율 (0~1 사이값) | Double |
| CASH\_ADVANCE\_FREQUENCY | 연중 현금서비스 구매 개월 수 비율 | Double |
| CASH\_ADVANCE\_TRX | 현금 서비스 구매 횟수 | Double |
| PURCHASES\_TRX | 구매 횟수 | Double |
| CREDIT\_LIMIT | 신용카드 한도 | Double |
| PAYMENTS | 지불 총액 | Double |
| MINIMUM\_PAYMENTS | 기한 내 최소 지불 금액 | Double |
| PRC\_FULL\_PAYMENT | 연중 기한 내 전액 지불 개월 수 비율 (0~1 사이값) | Double |
| TENURE | 신용카드 서비스 이용기간 | Double |

- 필요 패키지/라이브러리 목록

|  |  |
| --- | --- |
| Brightics |  |
| R | dplyr, data.table, tidyr, cluster, tree |
| Python | import pandas as pd  import numpy as np  from sklearn.metrics import silhouette\_samples, silhouette\_score  from sklearn.cluster import KMeans  from sklearn.tree import DecisionTreeRegressor |

분석을 수행하기 전, 상기 데이터를 이용하여 아래의 전처리를 수행하시오.

단계 1: ‘신용카드 한도(CREDIT\_LIMIT)’와 ‘기한 내 최소 지불 금액(MINIMUM\_PAYMENTS)’의 결측 값(Null)을 각 컬럼의 평균값으로 대체하시오.

상기 전처리를 완료한 데이터셋(데이터셋명: **card1**)을 이용하여 다음 1~3번 문제에 답하시오.

**1**. **(card1을 이용하여)** ’연간 평균 잔고액(BALANCE)’이 많을수록, 그리고 ‘신용카드 서비스 이용기간(TENURE)’이 길수록 ‘신용카드 한도(CREDIT\_LIMIT)’ 역시 높을 것으로 예상해볼 수 있다. 이들 변수 간의 관계를 파악하여, 추후 고객의 신용카드 한도 조정에 근거 자료로 활용하고자 한다.

‘신용 카드 서비스 이용기간(TENURE)’ 별로 ‘연간 평균 잔고액(BALANCE)’과 ‘신용카드 한도(CREDIT\_LIMIT)’ 간 피어슨(Pearson) 상관 계수를 계산하고, 이 중 가장 큰 값을 구하시오.

* 소수점 셋째 자리에서 반올림하여 소수점 둘째 자리까지 기술하시오.

**(답안예시) 0.12**

**2.** **(card1을 이용하여)** 신용카드 판매 전략을 수립하기 위해 고객 세분화(Customer Segmentation) 를 수행할 수 있다. 일시불 구매 금액이 높은 고객 Segment를 도출하기 위해 다음 단계에 따라 분석을 수행하고 질문에 답하시오.

단계 1: ‘고객 ID(CUST\_ID)’를 제외한 모든 변수(17개)에 대해 Z-score 표준화(Standardization) 한다.

단계 2: 표준화된 변수들에 대해 K-means 군집 분석을 수행한다. 이 때, 군집 수는 2~5개 중 K-means Silhouette 를 통해 구한 최적의 K로 설정한다.

단계 3: 단계 2에서 도출한 각 군집 별로 ‘일시불 구매 총액(ONEOFF\_PURCHASES)’의 평균을 계산한다.

|  |  |
| --- | --- |
| Brightics | Seed=1234  문제 지시 외 Default 값 사용 |
| R | library(cluster)  set.seed(12345)  표준화 : scale() 함수의 center=T, scale=T 옵션 사용  Silhouette : silhouette() 함수의 sil\_width 평균값 기준  문제 지시 외 Default 값 사용 |
| Python | from sklearn.metrics import silhouette\_samples, silhouette\_score  from sklearn.cluster import KMeans  random\_state=1234  문제 지시 외 Default 값 사용 |

군집 별 ‘일시불 구매 총액(ONEOFF\_PURCHASES)’의 평균 중 가장 큰 값을 구하시오.

* 소수점 셋째 자리에서 반올림하여 소수점 둘째 자리까지 기술하시오.

**(답안예시) 1200.34**

**3. (card1을 이용하여)** 이번에는 ‘일시불 구매 총액(ONEOFF\_PURCHASES)’ 예측 모델을 Target Marketing에 활용하고자 한다. 다음 단계에 따라 분석을 수행하고 질문에 답하시오.

단계 1: ‘고객 ID(CUST\_ID)’가 4의 배수가 아닌 데이터를 Train Set으로, 4의 배수인 데이터를 Test Set으로 분할한다.

단계 2: Train Set으로 아래 조건에 따라 의사결정나무 회귀모델을 학습한다.

* 독립 변수(총 16개): ‘고객 ID(CUST\_ID)’, ‘일시불 구매 총액(ONEOFF\_PURCHASES)’을 제외한 모든 컬럼
* 종속 변수: ‘일시불 구매 총액(ONEOFF\_PURCHASES)’
* 툴별 가이드

|  |  |
| --- | --- |
| Brightics | Seed=1234  문제 지시 외 Default 값 사용 |
| R | set.seed(1234)  library(tree)  Decision Tree Regression : tree() 함수 사용  문제 지시 외 Default 값 사용 |
| Python | from sklearn.tree import DecisionTreeRegressor  random\_state=1234  문제에서 지시한 것 외에는 Default 값 사용 |

단계 3: 생성된 모델을 Test Set에 적용하여 ‘일시불 구매 총액(ONEOFF\_PURCHASES)’을 예측한다.

단계 3에서 얻은 예측 결과를 평가하기 위해, 아래 정의된 Measure 를 구하시오.

: 예측값, : 실제값

* 는 소수점 둘째 자리에서 반올림하여 소수점 첫째 자리까지 기술하시오.

**(답안예시) 1200.3**

**□ 대문항 세트2**

교육 전문기관 분석팀에서는 교육 수강자의 정보를 바탕으로 진로설계 프로그램을 운영하기 위해 다음 정보를 수집하였다.

**DS\_Sample\_2.csv** (구분자: comma(“,”), 19,158 Rows, 15 Columns, UTF-8 인코딩)

|  |  |  |
| --- | --- | --- |
| **컬럼** | **정의** | **Type** |
| enrollee\_id | 수료자 ID | Double |
| city | 도시 코드 | String |
| city\_development\_index | 도시 발전 지표 | Double |
| gender | 성별 | String |
| relevant\_experience | 관련 분야 경험 여부 | String |
| enrolled\_university | 수강 과목명 | String |
| education\_level | 학력 | String |
| major\_discipline | 전공 | String |
| experience | 경력 | String |
| company­\_size | 현 직장 직원 수 | String |
| company\_type | 현 직장 유형 | String |
| last\_new\_job | 전 직장 근속연수 | String |
| training\_hours | 수료 시간 | Double |
| target | 이직 희망 여부 (0: 비희망, 1: 희망) | Double |
| Xgrp | Train/Test Set 구분 | String |

- 필요 패키지/라이브러리 목록

|  |  |
| --- | --- |
| Brightics |  |
| R | data.table, dplyr, tidyr, class |
| Python | Import pandas as pd  Import numpy as np  from sklearn.linear\_model import LogisticRegression  from sklearn.neighbors import KNeighborsClassifier |

분석을 수행하기 전, 상기 데이터를 이용하여 아래의 전처리를 수행하시오.

단계 1: 분석에 사용하지 않을 city, company\_size, company\_type 컬럼을 제거하시오.

단계 2: 각 문자형(String Type) 컬럼에 결측치(null/empty space)가 하나라도 존재하는 행(row)은 모두 제거하시오.

단계 3: experience 컬럼의 값이 ‘>20’ 또는 ‘<1’인 값을 제거하고 experience 컬럼의 타입을 정수형(Integer)으로 변환하시오.

단계 4: last\_new\_job 컬럼의 값이 ‘>4’ 또는 ‘never’인 값을 제거하고 last\_new\_job컬럼의 타입을 정수형(Integer)으로 변환하시오.

상기 전처리를 완료한 데이터셋(데이터셋명: **job2**, 7,522 Rows)을 이용하여 다음 4~6번 문제에 답하시오.

**4**. **(job2를 이용하여)** ‘관련 분야 경험 여부(relevant\_experience)’에 따른 ‘이직 희망 여부(target)’를 기술통계량으로 확인하고자 한다.

관련 분야 경험이 없는(relevant\_experience=‘No relevant experience’) 수료자 중 이직을 희망(target=’1’)하는 수료자의 비율을 A, 관련 분야 경험이 있는(relevant\_experience=’Has relevant experience’) 수료자 중 이직을 희망(target=’1’)하는 수료자의 비율을 B라 할 때, A/B를 구하시오.

* 소수점 셋째 자리에서 반올림하여 소수점 둘째 자리까지기술하시오.

**(답안예시) 12.34**

**5. (job2를 이용하여)** ‘이직 희망 여부(target)’에 영향을 주는 변수들을 확인하고자 한다. 다음 절차에 따라 로지스틱 회귀분석(Logistic Regression)을 수행하고 질문에 답하시오.

단계 1: gender, relevant\_experience, enrolled\_university, education\_level, major\_discipline 변수로부터 더미 변수들을 생성한다. 단, 각 변수로부터 더미 변수를 생성할 때 마지막으로 등장하는 범주는 제외하도록 한다.

(여기서 마지막으로 등장하는 범주란, 각 컬럼의 값을 사전 순으로 나열하였을 때 마지막으로 등장하는 값이다. 예를 들어, ‘col’ 변수의 범주가 [‘A’,’C’,’B’,’B’,’C’,’A’]의 값을 가진다면 사전 상의 마지막 값인 C가 제외된다)

단계 2: 단계 1에서 생성한 더미 변수와 city\_development\_index, experience, last\_new\_job, training\_hours, target, Xgrp 변수를 결합하여 새로운 데이터셋(데이터셋명: **job2\_2**, 이 데이터셋은 문제 6에서도 활용)을 구성한다. 이 때, target, Xgrp를 제외한 데이터셋의 컬럼은 아래 순서에 따르도록 한다.

- city\_development\_index

- experience

- last\_new\_job

- training\_hours

- gender의 더미 변수

- relevant\_experience의 더미 변수

- enrolled\_university의 더미 변수

- education\_level의 더미 변수

- major\_discipline의 더미 변수

단계 3. 단계 2에서 구성한 데이터셋 **job2\_2**로 다음 조건에 따라 상수항(Intercept)이 포함된 로지스틱 회귀분석을 수행한다.

* 종속 변수 : target
* 독립 변수(총 16개) : target과 Xgrp를 제외한 나머지 변수
* 회귀식에 포함되는 독립 변수의 순서를 컬럼의 순서와 일치시킨다.

|  |  |
| --- | --- |
| Brightics | Inverse of Regularization=100000  Seed=1234  문제 지시 외 Default 값 사용 |
| R | glm()  문제 지시 외 Default 값 사용 |
| Python | import pandas as pd  pd.get\_dummies() 사용  from sklearn.linear\_model import LogisticRegression  C=100000, random\_state=1234  문제 지시 외 Default 값 사용 |

상수항을 제외한 나머지 변수들에 대한 Odds Ratio중 가장 큰 값을 기술하시오.

의 Odds Ratio

* 소수점 셋째 자리에서 버림하여 둘째 자리까지 기술하시오.

**(답안예시) 12.34**

**6. (job2\_2를 이용하여)** 전체 데이터를 Train과 Test Set으로 나누고, Train Set으로 학습한 모델을 Test Set에 적용하여 모델을 평가하고자 한다. 다음 절차에 따라 분석을 수행하고 질문에 답하시오.

단계 1: 문제 5번 2단계에서 구성한 데이터셋 job2\_2에서Xgrp 컬럼의 값이 ‘train’인 경우 Train Set으로, ‘test’인 경우 Test Set으로 정의하여 분할한다.

단계 2: 아래 가이드에 따라 Train Set으로 K-NN 분류 모델을 학습하고, 이 모델을 Test Set에 적용한다.

* 종속 변수: 이직 희망 여부(target)
* 독립 변수(총 16개): 이직 희망 여부(target)와 Train/Test set 구분 변수(Xgrp)를 제외한 모든 변수
* Euclidean 거리 기준 가장 가까운 5개 데이터의 ‘이직 희망 여부(target)’를 활용하여 예측

|  |  |
| --- | --- |
| Brightics | 문제에서 지시한 것 외에는 Default 값 사용 |
| R | set.seed(1234)  library(class)  KNN : knn() 함수 사용  문제에서 지시한 것 외에는 Default 값 사용 |
| Python | from sklearn.neighbors import KNeighborsClassifier  문제 지시 외 Default 값 사용 |

단계 3: 예측 결과를 바탕으로 아래 정의된 지표 *A*를 계산하여 기술하시오.

* 는 소수점 셋째 자리에서 반올림하여 소수점 둘째 자리까지 기술하시오.

**(답안예시) 0.12**

**□ 대문항 세트3**

영화 스트리밍 사이트 운영진은 등록된 영화들에 대한 데이터를 분석하여, 개봉 년도, 평점, 투표 수, 감독 등 변수들 간 관계에 대해 알아 보고자 한다.

**DS\_Sample\_3.csv** (구분자: comma(“,”), 188 Rows, 5 Columns, UTF-8 인코딩)

|  |  |  |
| --- | --- | --- |
| **컬럼** | **정의** | **Type** |
| Title | 영화 제목 | String |
| AirDate | 영화 공개 날짜 | String |
| Rating | 평점 | Double |
| Num\_Votes | 투표 수(평점을 남긴 수) | Double |
| DirectedBy | 감독 | String |

- 필요 패키지/라이브러리 목록

|  |  |
| --- | --- |
| Brightics |  |
| R | data.table, dplyr, tidyr |
| Python | import pandas as pd  import numpy as np  import datetime  from statsmodels.stats.anova import anova\_lm  from statsmodels.formula.api import ols  from sklearn.linear\_model import LinearRegression |

분석을 수행하기 전, 상기 데이터를 이용하여 아래의 전처리를 수행하시오.

단계 1: 영화 공개 날짜(AirDate)의 연도 정보를 기준으로, 2005년에서 2007년은 ‘A’, 2008년에서 2010년은 ‘B’, 2011년에서 2013년은 ‘C’ 값을 가지는 공개년도그룹(변수명: group) 변수를 생성하시오.

상기 전처리를 완료한 데이터셋(데이터셋명: **movie3**)을 이용하여 다음 7~9번 문제에 답하시오.

**7. (movie3을 이용하여)** 감독 별로 영화 흥행 결과를 알아보고자 한다. 여기서 영화의 흥행 변수 ‘success’는 다음과 같이 정의된다.

‘감독(DirectedBy)’ 별로 ‘흥행(success)‘ 변수의 평균을 계산하고, 이 중 가장 큰 값을 소수 부분을 버리고 정수 부분만 기술하시오.

**(답안예시)** 1234

**8. (movie3을 이용하여)** 영화 평가에 참여한 총 투표 규모는 영화의 컨텐츠 뿐만 아니라 서비스 공개 시기에 따라 달라질 가능성이 있다. 영화 공개 년도에 따라 투표 수에 차이가 존재하는지 검정하기 위해 수립한 가설은 다음과 같다.

대립 가설(H1): 공개년도그룹(group)에 따라 평균 투표 수(Num\_Votes)는 달라진다.

적절한 검정 방법을 택하여 위 가설을 검정할 때, 검정통계량 값(A)과 요인의 자유도(B)의 합(A+B)을 기술하시오.

* 그룹 간 등분산성 및 정규성은 만족한다고 가정한다.
* 툴별 가이드

|  |  |
| --- | --- |
| Brightics | 문제 지시 외 Default 값 사용 |
| R | aov() 함수 사용  문제 지시 외 Default 값 사용 |
| Python | from statsmodels.stats.anova import anova\_lm  from statsmodels.formula.api import ols  문제 지시 외 Default 값 사용 |

소수점 셋째 자리에서 반올림하여 소수점 둘째 자리까지 기술하시오.

**(답안예시) 12.34**

**9. (movie3을 이용하여)** 행동 경제학에서는 개인의 의사와 무관하게 집단 전체가 동일한 방향으로 움직이는 군집 행동(Herding Behavior)이 존재한다고 본다. 이 이론에 따르면 특정 영화 관람객이 많아지고 대중적 인기를 끌수록 그 영화에 대한 평균 평점은 계속해서 높아질 가능성이 있다. 이 현상이 실제로 나타나는지 알아보기 위해 ‘투표 수(Num\_Votes)’와 ‘평점(Rating)’ 간 관계를 분석하고자 한다.

아래 가이드에 따라 OLS(Ordinary Least Squares) 방식의 선형회귀 모델링을 수행하고 질문에 답하시오.

- 종속 변수: 평점(Rating)

- 독립 변수(총 3개): 공개년도그룹(group)를 one-hot encoding한 결과 컬럼(2개), 투표 수(Num\_Votes)

- 회귀식에 상수항을 포함할 것

|  |  |
| --- | --- |
| Brightics | 문제 지시 외 Default 값 사용 |
| R | 문제에서 지시한 것 외에는 Default 값 사용 |
| Python | from sklearn.linear\_model import LinearRegression  문제 지시 외 Default 값 사용 |

추정된 모델을 기반으로, 투표 수(Num\_Votes)가 5000, 공개년도그룹(group)이 C인 경우 평점(Rating)의 예측 값을 기술하시오.

* 소수점 셋째 자리에서 반올림하여 소수점 둘째 자리까지 기술하시오.

**(답안예시) 12.34**

**실기툴별 정답**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **배점** | **Brightics** | **Python** | **R** |
| **1** | 10 | 0.95 | 0.95 | 0.95 |
| **2** | 10 | 3946.19 | 3946.19 | 3946.19 |
| **3** | 15 | 2383.8 | 2383.8 | 1599.4 |
| **4** | 10 | 1.77 | 1.77 | 1.77 |
| **5** | 10 | 1.67 | 1.67 | 2.22 |
| **6** | 15 | 0.71 | 0.71 | 0.71 |
| **7** | 10 | 32720 | 32720 | 32720 |
| **8** | 10 | 18.03 | 18.03 | 18.03 |
| **9** | 10 | 9.03 | 9.03 | 9.03 |

\* 시험환경에서 실행한 정답 기준으로 작성하였으며, 문항별 정답은 사용 라이브러리/패키지의 버전 차이로 달라질 수 있습니다.