

부록

빅데이터 분석 기사 기출 실습

[수업 목표]

이번 시간에는 빅데이터 분석기사 실기 시험을 같이 준비해보겠습니다

[수업 개요]

빅데이터 분석기사 응시환경 접속 방법

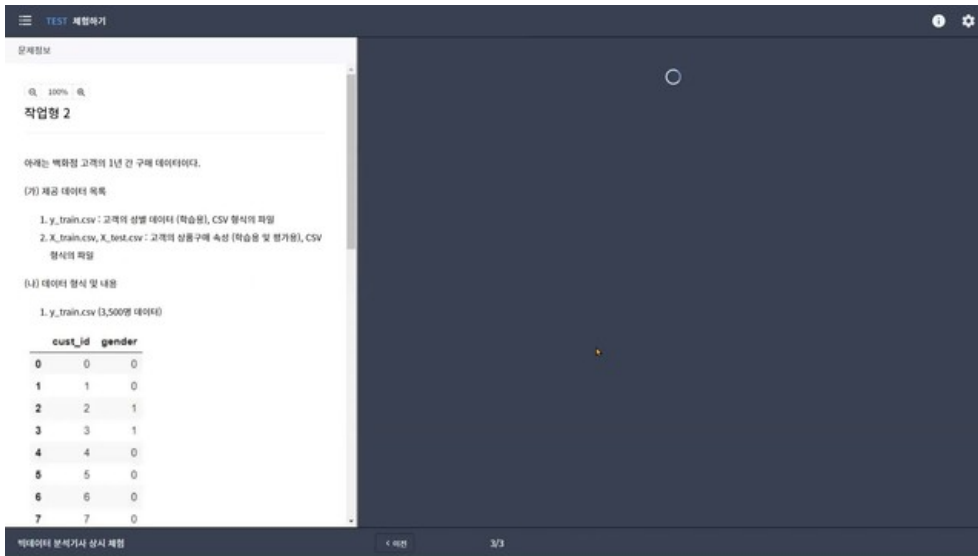
0:15



[작업형 제 2유형 풀이 시작]

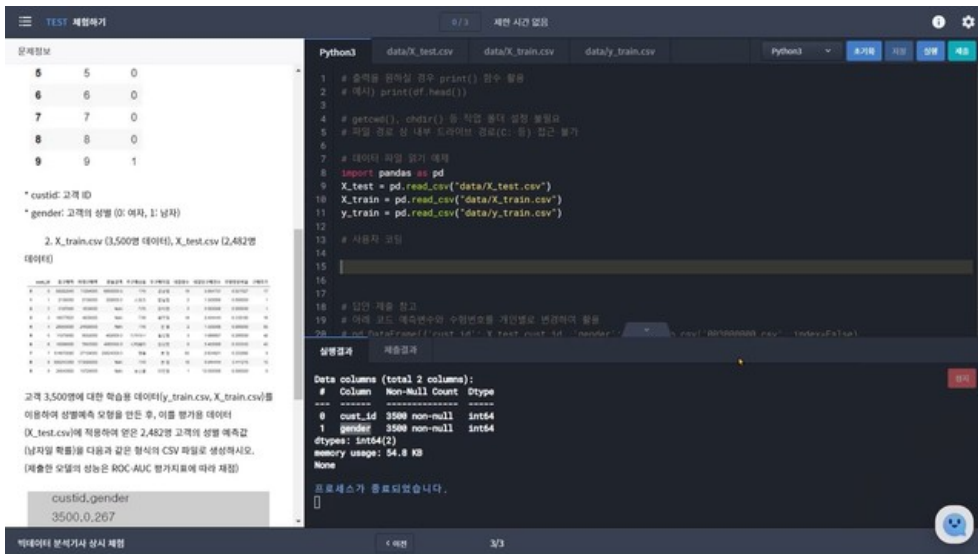
1:42

머신러닝 모델링과 평가지표를 사용하실 수 있다면 충분히 풀실 수 있는 문제가 출제됩니다.



[데이터 전처리]

09:13



[데이터 모델링]

sklearn 의 randomforest classifier를 사용하여 train데이터를 학습합니다.

17:50

문제정보

5	5	0
6	6	0
7	7	0
8	8	0
9	9	1

* custid: 고객 ID
* gender: 고객의 성별 (0: 여자, 1: 남자)

2. X_train.csv (3,500명 데이터), X_test.csv (2,482명 데이터)

고객 3,500명에 대한 학습용 데이터(y_train.csv, X_train.csv)를 이용하여 성능평가 요청을 받은 후, 이를 평가용 데이터 (X_test.csv)에 적용하여 얻은 2,482명 고객의 성별 예측값 (남자일 확률)을 다음과 같은 형식의 CSV 파일로 생성하시오. (제출한 모델의 성능은 ROC-AUC 평가치로 평가)

custid,gender
3500,0.267

```
Python3
data/X_test.csv data/X_train.csv data/y_train.csv
Python3
14
15 # 1) y_train, set_index
16 # 2) X_train, X_test, split 지점
17 # 3) X_train, X_test, split 지점
18
19 y_train = y_train.set_index('cust_id')
20 y_train = y_train['gender']
21
22 X_train['성별'] = X_train['성별'].fillna(0)
23 X_test['성별'] = X_test['성별'].fillna(0)
24
25 X_train = pd.get_dummies(X_train, columns = ['주구매상품', '주구매지점'])
26 X_test = pd.get_dummies(X_test, columns = ['주구매상품', '주구매지점'])
27 X_test['주구매상품_소형가전'] = 0
28
29
30
31 # 일단 예측 결과
32 # 이리 코드 계속 반복해서 주입변수, 개인별로 변경하여, 활용
33 # pd.DataFrame({'cust_id': X_test.cust_id, 'gender': pred}).to_csv('ROC3000000.csv', index=False)
```

실행결과

```
프로세스가 시작되었습니다. (입력값을 직접 입력해 주세요)
> [3500, 74] (2482, 74)
Index([], dtype='object')
프로세스가 종료되었습니다.
```

[데이터 평가]

from sklearn.metrics import roc_auc_score를 사용하여 분류 모델을 평가해줍니다.

22:33

문제정보

5	5	0
6	6	0
7	7	0
8	8	0
9	9	1

* custid: 고객 ID
* gender: 고객의 성별 (0: 여자, 1: 남자)

2. X_train.csv (3,500명 데이터), X_test.csv (2,482명 데이터)

고객 3,500명에 대한 학습용 데이터(y_train.csv, X_train.csv)를 이용하여 성능평가 요청을 받은 후, 이를 평가용 데이터 (X_test.csv)에 적용하여 얻은 2,482명 고객의 성별 예측값 (남자일 확률)을 다음과 같은 형식의 CSV 파일로 생성하시오. (제출한 모델의 성능은 ROC-AUC 평가치로 평가)

custid,gender
3500,0.267

```
Python3
data/X_test.csv data/X_train.csv data/y_train.csv
Python3
24
25 X_train['성별'] = X_train['성별'].fillna(0)
26 X_test['성별'] = X_test['성별'].fillna(0)
27
28 X_train = pd.get_dummies(X_train, columns = ['주구매상품', '주구매지점'])
29 X_test = pd.get_dummies(X_test, columns = ['주구매상품', '주구매지점'])
30 X_test['주구매상품_소형가전'] = 0
31
32 from sklearn.model_selection import train_test_split
33 X_train_dum, X_val, y_train_dum, y_val = train_test_split(X_train, y_train, stratify = y_train, test_size=0.2)
34
35 from sklearn.ensemble import RandomForestClassifier
36 rf = RandomForestClassifier()
37 rf.fit(X_train_dum, y_train_dum)
38
39
40
41
42 # 일단 예측 결과
43 # 이리 코드 계속 반복해서 주입변수, 개인별로 변경하여, 활용
```

실행결과

```
프로세스가 시작되었습니다. (입력값을 직접 입력해 주세요)
>
cust_id
2874 171699400 17535000 38440000.0 ...
3282 28273000 15295000 0.0 ...
3189 635300 1750000 0.0 ...
545 49532540 14000000 0.0 ...
2588 78711950 13910000 0.0 ...
...
2588 166000 181200 0.0 ...
367 61699218 15887000 0.0 ...
```

[모델 튜닝 및 평가]

정확도를 향상시키기 위해 매개변수를 최적화 시켜주어야 합니다.

27:15

TEST 세팅하기

0/3 재전 시간 없음

문제정보

5 5 0

6 6 0

7 7 0

8 8 0

9 9 1

* custid: 고객 ID

* gender: 고객의 성별 (0: 여자, 1: 남자)

2. X_train.csv (3,500명 데이터), X_test.csv (2,482명 데이터)

customer_id, gender, product_id, product_name, price, quantity, total_price, discount, rating, review_text

1, 0, 1000001, '스마트폰', 1000, 1, 1000000, 0.0, 4.5, '정말 좋아요'

2, 1, 1000002, '노트북', 2000, 1, 2000000, 0.0, 4.2, '화면이 크고 성능이 좋아요'

3, 0, 1000003, '스마트폰', 800, 2, 1600000, 0.0, 4.8, '가격이 저렴하고 성능이 좋아요'

4, 1, 1000004, '노트북', 1500, 1, 1500000, 0.0, 4.1, '디자인이 예뻐요'

5, 0, 1000005, '스마트폰', 1200, 1, 1200000, 0.0, 4.6, '카메라가 좋아요'

6, 1, 1000006, '노트북', 1800, 1, 1800000, 0.0, 4.3, '배터리가 오래가요'

7, 0, 1000007, '스마트폰', 900, 1, 900000, 0.0, 4.7, '화면이 선명해요'

8, 1, 1000008, '노트북', 1600, 1, 1600000, 0.0, 4.4, '키보드 타감이 좋아요'

9, 0, 1000009, '스마트폰', 1100, 1, 1100000, 0.0, 4.9, '화면이 넓어요'

10, 1, 1000010, '노트북', 1700, 1, 1700000, 0.0, 4.0, '성능이 좋아요'

고객 3,500명에 대한 학습용 데이터(y_train.csv, X_train.csv)를 이용하여 성별 예측 모델을 만든 후, 이를 평가용 데이터 (X_test.csv)에 적용하여 얻은 2,482명 고객의 성별 예측값 (남자일 확률)을 다음과 같은 형식의 CSV 파일로 생성하시오. (제출한 모델의 성능은 ROC-AUC 평가치로 따라 채점)

custid,gender

3500,0.267

Python3

data/X_test.csv data/X_train.csv data/y_train.csv

Python3 초기화 실행 새로고침

```

23 X_test = X_test.set_index('cust_id')
24
25 X_train['완충금액'] = X_train['완충금액'].fillna(0)
26 X_test['완충금액'] = X_test['완충금액'].fillna(0)
27
28 X_train = pd.get_dummies(X_train, columns = ['주구매상품', '주구매지점'])
29 X_test = pd.get_dummies(X_test, columns = ['주구매상품', '주구매지점'])
30 X_test['주구매상품_소형가전'] = 0
31
32 from sklearn.model_selection import train_test_split
33 X_train_dum, X_val, y_train_dum, y_val = train_test_split(X_train, y_train, stratify = y_train, test_size=0.2)
34
35
36 from sklearn.ensemble import RandomForestClassifier
37 rf = RandomForestClassifier()
38 rf.fit(X_train_dum, y_train_dum)
39
40
41 from sklearn.metrics import roc_auc_score
42 print(roc_auc_score(y_val, rf.predict_proba(X_val)[:,1]))
43
44
45 # 중간 결과 출력
46 # 아래 코드는 예측결과를 수정번호를 개인별로 변경하여 출력
47 # pd.DataFrame({'cust_id': X_test.cust_id, 'gender': pred}).to_csv('003000000.csv', index=False)

```

실행결과 재실행결과

프로세스가 시작되었습니다. (입력값을 직접 입력해 주세요)

> 0.6279914985465818

프로세스가 종료되었습니다.

[다음 수업 예고]

다음 시간에는 ADP 시간분배에 대해 다뤄보겠습니다

CLASS101

본 수업 자료를 무단 복제, 가공 및 배포시에 저작권 침해로 법적 책임을 물을 수 있습니다.
a2FrYW8xNDkwMTMyMjQ0