💯

작업형 2 기출 분석

5회 : 회귀 (RMSE)

[가격 예측] 중고 자동차

- 자동차 가격을 예측해주세요!
- 예측할 값(y): price
- 평가: RMSE (Root Mean Squared Error)
- data: train.csv, test.csv
- 제출 형식: result.csv파일을 아래와 같은 형식(수치형)으로 제출

```
id,price
0,11000
1,20500
2,19610
...
1616,11995
```

4회 : 다중분류 - 예측 '값', Macro f1-score

[마케팅] 자동차 시장 세분화

- 자동차 회사는 새로운 전략을 수립하기 위해 4개의 시장으로 세분화했습니다.
- 기존 고객 분류 자료를 바탕으로 신규 고객이 어떤 분류에 속할지 예측해주세요!
- 예측할 값(y): "Segmentation" (1,2,3,4)
- 평가: Macro f1-score
- data: train.csv, test.csv
- 제출 형식

```
ID,Segmentation
458989,1
458994,2
459000,3
459003,4
```

- 아래 코드 예측변수와 수험번호를 개인별로 변경하여 활용
- pd.DataFrame({'ID': test.ID, 'Segmentation': pred}).to_csv('003000000.csv', index=False)

3회 : 이진분류 - 예측 ‘확률’ (roc_auc_score)

여행 보험 패키지 상품을 ‘구매할’ 확률 값 (클래스 1 예측확률)

- 예측할 값(y): TravelInsurance (여행보험 패지지를 구매 했는지 여부 0:구매안함, 1:구매)
- 평가: roc-auc 평가지표
- data: t2-1-train.csv, t2-1-test.csv
- 제출 형식

```
id,TravelInsurance
0,0.3
1,0.48
2,0.3
3,0.83
```

⭐ **문제 분석**

- 특이점 : train, test의 범주형 칼럼 nunique가 달랐음
- train에 없는 범주가 test에는 있어서, 인코딩 시 에러 남
  - 따라서 인코딩 전에 train, test 데이터를 합쳐서 인코딩 후 다시 분리해 주기

```
# train data
Employment Type GraduateOrNot FrequentFlyer  \
count                          1490          1490          1490
unique                            2             2             2
top     Private Sector/Self Employed           Yes            No
freq                           1056          1270          1175

       EverTravelledAbroad
count                 1490
unique                   2
top                     No
freq                  1209

# test data
                    Employment Type GraduateOrNot FrequentFlyer  \
count                          497           497           497
unique                           3             2             2
top     Private Sector/Self Employed           Yes            No
freq                           360           422           395

       EverTravelledAbroad
count                  497
unique                   2
top                     No
freq                   398
```

2회 : 분류 - 예측 '확률' (roc_auc_score)

전자상거래 배송 데이터

제품 배송 시간에 맞춰 배송되었는지 예측모델 만들기

학습용 데이터 (X_train, y_train)을 이용하여 배송 예측 모형을 만든 후, 이를 평가용 데이터 (X_test)에 적용하여 얻은 예측 확률값을 다음과 같은 형식의 CSV파일로 생성하시오(제출

한 모델의 성능은 ROC-AUC 평가지표에 따라 채점)

- 제출형식

```
ID, Reached.on.Time_Y.N
4733,0.6
2040,0.8
5114,0.45
2361,0.23
5996,0.43
```

공식예제 : 분류 - 예측 '확률' (roc_auc_score)

백화점 고객의 1년 간 구매 데이터

- y_train.csv : 학습용 고객 성별 데이터 (0:여자, 1:남자)
- X_train.csv, X_test.csv : 학습, 평가용 고객의 상품구매 속성
- train으로 성별예측 모형을 만든 후, test 데이터에 적용하여 고객의 성별 예측값(남자일 확률)을 csv 결과 파일로 제출
- 평가지표 : roc_auc

아래와 같은 형태의 csv 결과 파일 생성

```
custid, gender
3500, 0.267
3501, 0.578
```

⭐ 문제 분석

- 특이점 : train, test의 범주형 칼럼 nunique가 달랐음
- train에 있는 범주가 test에는 없어서, 인코딩 시 에러가 쉽게 남
  - 라벨인코딩 적용

```
# train
       주구매상품 주구매지점
count   3500  3500
unique    42    24
top       기타  본  점
freq     595  1077

# test
       주구매상품 주구매지점
count   2482  2482
unique    41    24
top       기타  본  점
freq     465   726

KeyError: "['주구매상품_소형가전'] not in index"
```