<FDS 프로젝트 모델 학습 오류 공유 보고서>

| **오류공유 보고서** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|
| 제목 | 모델 학습 중, 학습 데이터 오류 | | | | | | | |
| 대상 | 전처리 학습 데이터 | | | | | | | |
| 발행일자 | 2024/09/10 | | | | | | | |
| 작성자 | 김동훈 | | | | | | | |
| **오 류 내 용** | | | | | | | | |
| 전처리된 학습 데이터를 활용한 모델 학습 중, 에러 발생  에러 코드:  Traceback (most recent call last):  File "c:\Users\USER\Desktop\fds데이터처리코드\FDS\_\_model\_isolationForest.py", line 64, in <module>  y\_pred = model.predict(x\_val)  ^^^^^^^^^^^^^^^^^^^^  File "C:\Users\USER\anaconda3\Lib\site-packages\sklearn\ensemble\\_iforest.py", line 369, in predict  decision\_func = self.decision\_function(X)  ^^^^^^^^^^^^^^^^^^^^^^^^^  File "C:\Users\USER\anaconda3\Lib\site-packages\sklearn\ensemble\\_iforest.py", line 404, in decision\_function  return self.score\_samples(X) - self.offset\_  ^^^^^^^^^^^^^^^^^^^^^  File "C:\Users\USER\anaconda3\Lib\site-packages\sklearn\ensemble\\_iforest.py", line 431, in score\_samples  X = self.\_validate\_data(X, accept\_sparse="csr", dtype=tree\_dtype, reset=False)  ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^  File "C:\Users\USER\anaconda3\Lib\site-packages\sklearn\base.py", line 608, in \_validate\_data  self.\_check\_feature\_names(X, reset=reset)  File "C:\Users\USER\anaconda3\Lib\site-packages\sklearn\base.py", line 535, in \_check\_feature\_names  raise ValueError(message)  ValueError: The feature names should match those that were passed during fit.  Feature names unseen at fit time:  - category  - city  - first  - last  - merchant  Feature names seen at fit time, yet now missing:  - category\_entertainment  - category\_food\_dining  - category\_gas\_transport  - category\_grocery\_pos  - category\_health\_fitness  - ... | | | | | | | | |
|
|
| **원인 요약** | | | | | | | | |
| 데이터 전처리 이후, 최종 데이터셋에 문자열 데이터(object)가 존재. 해당 데이터들은 전처리 중, ‘법주형 변수 처리’ 및 ‘데이터 변환 및 인코딩’ 과정에서 수치형으로 변환할 수 있으나, 데이터셋에 무의미한 데이터 및 컬럼이 대량으로 추가되는 문제가 발생하여 해당 과정을 테스트만 한 이후 최종 데이터셋 저장 과정에서는 진행하지 않음. 이후, 모델 학습 과정에서 해당 문자열 데이터를 처리하는 과정이 ‘x\_train’에서만 적용되어 ‘예측 수행 및 성능 평과’ 과정에서 사용된 ‘x\_val’과 ‘y\_val’ 데이터셋 컬럼의 차이로 인하여 에러가 발생함. | | | | | | | | |
|
|
|
|
| **시 정 조 치** | | | | | | | | |
| 1. 데이터 전처리 과정에서 범주형 데이터를 수치형 데이터로 변환 후, 최종 데이터셋 수정. (해당 과정을 거치면 무의미한 데이터와 컬럼이 대량으로 추가 됨.) 2. 추가적 데이터 처리    1. 최종 데이터셋에서 문자열 데이터(object) 컬럼 제거    2. 모든 최종 데이터셋의 문자열 데이터 처리. (One-Hot Encoding 방식 적용)    3. 고객 이름(first, last) 컬럼의 경우, 데이터 형태는 문자열(object)이지만 개인정보 보호를 위해 임의의 고객 ID를 부여 하였으므로 제거한다. | | | | | | | | |
|
|

| **오류공유 보고서** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|
| 제목 | 모델 학습 중, 하이퍼파라미터 튜닝 에러(1) | | | | | | | |
| 대상 | 하이퍼파라미터, 그리드 서치 | | | | | | | |
| 발행일자 | 2024/09/10 | | | | | | | |
| 작성자 | 김동훈 | | | | | | | |
| **오 류 내 용** | | | | | | | | |
| 하이퍼파라미터 튜닝 과정 중, 경고 메시지 발생.  에러 코드:  warnings.warn( C:\Users\USER\anaconda3\Lib\site-packages\sklearn\model\_selection\\_validation.py:993: UserWarning: Scoring failed. The score on this train-test partition for these parameters will be set to nan. Details: | | | | | | | | |
| **원인 요약** | | | | | | | | |
| GridSearchCV 실행 중 특정 파라미터에 대해 평가 점수를 계산할 수 없어, 해당 파라미터 조합에 대한 점수를 nan으로 설정 한다.  주요 원인:   1. 잘못된 평가 지표: scoring 매개변수에 지정된 평가 지표가 현재 문제 유형이나 모델에 적합하지 않을 수 있다. 2. 데이터 형태 문제: 학습 데이터 또는 검증 데이터의 형태가 모델이 요구하는 형태와 다를 수 있다. 3. 샘플 수 부족: contamination 값이 너무 높아서 이상치로 분류된 샘플이 너무 적거나, 특정 클래스의 샘플이 너무 적어서 평가 지표를 계산할 수 없는 경우일 수 있다. 4. 모델 자체의 문제: 모델 구현에 오류가 있거나, 특정 하이퍼파라미터 조합에서 모델이 제대로 작동하지 않는 경우 오류가 발생할 수 있다. | | | | | | | | |
|
|
|
|
| **시 정 조 치** | | | | | | | | |
| 1. 평가 지표 확인: scoring 매개변수에 지정된 평가 지표가 현재 문제 유형과 모델에 적합한지 확인한다. IsolationForest 모델의 경우, roc\_auc 외에도 f1, precision, recall 등을 사용할 수 있다. 2. 데이터 형테 확인: x\_train과 x\_val의 형태가 모델이 요구하는 형태와 일치하는지 확인한다. 필요하다면 reshape 메서드 등을 사용하여 데이터 형태를 변환한다. 3. contamination 값 조정: contamination 값을 낮춰서 이상치로 분류되는 샘플 수를 늘린다. 4. 다른 하이퍼파라미터 조합 시도: param\_grid에 지정된 하이퍼파라미터 범위를 조정하여 다른 조합을 시도한다. 5. 모델 코드 검토: 모델 구현에 오류가 없는지 확인한다. 필요하다면 다른 라이브러리에서 제공하는 IsolationForest 모델을 사용한다. | | | | | | | | |
|
|

| **오류공유 보고서** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|
| 제목 | 모델 학습 중, 하이퍼파라미터 튜닝 에러(2) | | | | | | | |
| 대상 | 하이퍼파라미터, 그리드 서치 | | | | | | | |
| 발행일자 | 2024/09/10 | | | | | | | |
| 작성자 | 김동훈 | | | | | | | |
| **오 류 내 용** | | | | | | | | |
| 하이퍼파라미터 튜닝 과정 중, TypeError 발생.  에러 코드:  Traceback (most recent call last):  File "C:\Users\USER\anaconda3\Lib\site-packages\sklearn\model\_selection\\_validation.py", line 980, in \_score  scores = scorer(estimator, X\_test, \*\*score\_params)  ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^  TypeError: \_BaseScorer.\_\_call\_\_() missing 1 required positional argument: 'y\_true' | | | | | | | | |
| **원인 요약** | | | | | | | | |
| scorer 함수가 호출되었을 때, y\_true가 누락되어 발생.  scoring 함수가 지도 학습에서 y\_true(실제 레이블)를 기대하지만, Isolation Forest 모델은 비지도 학습이기 때문에 레이블이 없다.  Isolation Forest 모델은 이상 탐지를 위한 비지도 학습 모델로, fit 함수에서 레이블(y) 없이 입력 데티어(x\_train)만을 사용하지만, GridSearchCV는 지도 학습 모델을 기준으로 설계되어, 기본적으로 y\_true를 필요로 하는 scoring 메트릭을 기대한다. | | | | | | | | |
|
|
|
|
| **시 정 조 치** | | | | | | | | |
| 1. scoring 메트릭을 비지도 학습용으로 적합하게 설정    1. Isolation Forest 모델은 기본적으로 ROC AUC를 사용할 수 없기 떄문에, fit에 사용할 수 있는 적절한 스코어링 함수로 neg\_mean\_squared\_error 또는 사용자 정의 스코어링 함수를 사용한다.    2. make\_scorer를 정의하여 scoring 매개변수에 전달한다. 2. GridSearchCV에 y\_true 값을 포함하는 데이터셋을 전달 3. scoring 매개변수에 콜러블 객체를 직접 전달 | | | | | | | | |
|
|