<모델 학습 실패 원인 보고서>

| 과제명 | FDS 모델 학습 | | |
| --- | --- | --- | --- |
| 책임자 | 김동훈 | 보고서 작성 일자 | 2024/09/10 |

| 평가 결과 및 사유 | 전처리된 최종 학습 데이터를 이용하여 IsolationForest 모델을 활용한 FDS 학습 결과, Precision 및 Recall 이 0이 나오며 경고 발생한다.  C:\Users\USER\anaconda3\Lib\site-packages\sklearn\metrics\\_classification.py:1509: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.  \_warn\_prf(average, modifier, f"{metric.capitalize()} is", len(result)) |
| --- | --- |
| 실패 사유 | 1. Precision 및 Recall이 잘 정의되어 있지 않음.    1. 모델이 특정 클래스에 대해 예측을 거의 하지 않거나, 실제로 특정 클래스에 대한 샘플이 거의 없을 때 발생. 2. 불균형 데이터 문제    1. support에서 클래스 -1과 1은 매우 적고, 클래스 0이 대부분의 데이터를 차지한다. 이는 데이터가 매우 불균형 하다는 것을 의미한다. 즉, 학습 데이터의 is\_fraud 데이터가 극단적으로 적다는 것을 의미한다. |
| 향후 계획 | 1. zero\_division 매개변수 설정    1. zero\_dicision 매개변수를 사용하여, Precision과 Recall을 -으로 나누는 경우를 처리 가능하다. (근본적 문제 해결은 불가하다.) 2. 데이터 불균형 해결    1. Undersampling. 정상 데이터를 줄여서 비정상 데이터와 균형을 맞춘다..    2. Oversampling. 비정상 데이터를 증가시켜서 데이터의 균형을 맞춘다..(SMOTE)    3. 비율 조정. Isolation Forest 모델의 contamination 매개변수를 사용하여 이상치의 비율을 조정한다. 3. 모델 성능 개선    1. Isolation Forest 모델이 현재 데이터에 적합하지 않을 수 있다는 가능성이 존재하므로, 다른 모델을 시도해 볼 수 있다. (RandomForest, XGBoost, Logistic Regression 등) |
| 비고 |  |