데이터 분석

데이터 셋 구성

센서 데이터

: 각종 공정 과정에서 발생한 데이터들

: 제품의 ID인 Set ID라는 컬럼이 존재

set ID컬럼을 이용하여 4가지 데이터를 병합 해야함.

(1) fill1 dispensing.xlsx

- Fill : 내측 도포
- LOT ID, Set ID : 제품 하나의 고유 넘버링
- 데이터 : 레진 토출 좌표, 레진 토출 속도, 레진 토출 시간, 레진 토출양, 레진 대기 좌표, 공정 소요시간, 노즐 클린 좌표, uv경화속도, uv경화시간, uv경화위치 ...

(2) fill2 dispensing.xlsx

(3) Dam dispensing.xlsx

- Dam : 외측 도포
- 데이터 : 레진 토출 좌표, 레진 토출 속도, 레진 토출 시간, 레진 토출양, 레진 대기 좌표, 공정 소요시간, 노즐 클린 좌표, uv경화속도, uv경화시간, uv경화위치 ...

(4) Auto clave.xlsx

📣 **[바뀐 데이터]**
- train.csv
기존에 4개의 엑셀 파일로 제공되던 공정 데이터는
Set ID를 기준으로 병합되었고, Set ID를 제거한 하나의 단일 파일로 구성하여 학습 데이터(train.csv)로 제공합니다.

- test.csv
train.csv와 같은 방식으로 구성한 Test 데이터는 기존에 부여된
Set ID를 랜덤하게 변경하여 제공합니다.

두 데이터 모두, Set ID 뿐 아니라, 시간 데이터로도 모든 데이터를 식별할 수 있기 때문에, 새로운 데이터셋을 이전의 Set ID와 매칭하여 추론하는 것을 방지하기 위해, 시간 데이터는 모두 삭제하였습니다.

이번 대회에서는 학습 데이터 내의 target을 제외한 모든 칼럼을 Feature로 활용할 수 있습니다.

다만 그 방식은 현업에 적용 가능한 방식이어야 하며, 제품의 불량여부를 추론하는 시점에 활용 가능한 정보여야 합니다.

학습용 제품 목록

- train.csv
  - 모델에 입력하는 값인 X변수는 train.csv파일 내의 칼럼 중 target칼럼을 제외한 모든 칼럼이 해당됨
  - 예측해야 하는 대상인 Y변수는 target이라는 칼럼으로 제품의 이상 여부를 나타냄.
    - AbNormal : 제품에 각종 이상이 있다는 의미
    - Normal : 제품이 정상이라는 의미

테스트용 데이터

테스트용 데이터는 아래의 제출용 제품의 이상 여부를 예측하기 위한 센서 데이터들이 포함되어 있다.

모든 칼럼이 학습용 데이터와 동일하지만, 아래 두가지 내용이 다르다.

- 제품의 id가 포함되어 있다. 이는 제출용 제품 목록의 이상 여부를 채점하기 위해 사용됨
- target 칼럼이 비어있다.

제출용 제품 목록

- submission.csv
  - 예측 후 제출해야 하는 제품 목록
  - 제출용 제품 목록 파일에는 정답이 되는 target이 없다.
  - 예측을 통해 target 칼럼을 채워 넣고 제출해라.

Random state

베이스 라인 코드에는 일정한 성능이 나오도록 seed에 해당하는 random_state라는 변수 있음.

해당 값은 기본 110으로 설정, 임의로 수정 가능

결과 지표

분류 지표인 F1 Score를 활용하자

데이터 셋은 AbNormal과 Nornal의 개수 차이가 심한 편으로, 예측해야 하는 대상인 AbNormal을 기준으로 F1 Score를 계산한다.

채점방법

`submission.csv` 파일을 읽고 저장해야 합니다. 읽은 파일을 위의 안내의 따라 값을 채워 넣은 후에 나온 테스트 데이터를 학습된 모델에 적용합니다.

모델에서 나온 예측 결과는 아래와 같은 형태의 CSV 파일로 저장하여야 합니다. 파일 이름은 `submission.csv`로 해야 합니다.

해당 모델이 얼마나 잘 작동하는지 통계적으로 확인해보는 단계

예측/실제	True	False
True	True Positive	False Positive
False	False Negative	True Negative

- **F1 Score 계산**

Recall(재현율) : 실제로 True인 데이터를 모델이 True라고 인식하는 데이터의 수

$$\frac{TruePositives}{TruePositives + FalseNegatives}$$

Precision(정밀도) : 모델이 True로 예측한 데이터 중 실제로 True인 데이터의 수이다.

$$\frac{TruePositives}{TruePositives + FalsePositives}$$

F1 Score : 정밀도(precision)와 재현율(recall)의 조화평균을 계산되는 지표

F1 Score는 정밀도와 재현율을 조합하여 하나의 통계치를 반환한다.

$$2 * \frac{Precision * Recall}{Precision + Recall}$$
