제3회 K-인공지능 제조데이터 분석 경진대회

샘플 불량률을 활용한 품질 예측 프레임워크

Team 리트라이 | 부도현(팀장), 변유정, 서성현

목차

01

문제 정의

1-1 분석 배경 p.3

1-2 분석 목표 p.5

02

분석 모델

2-1 데이터 처리 과정 p.7

2-2 분석 모델 구축 p.9

03

분석 결과

3-1 분석 모델 학습 결과 p.11

04

결론

4-1 종합 결론 p.12

4-2 기대효과 p.13

☑ 제조 현장의 이슈: 데이터 품질 관리

데이터 품질 관리 체계의 부재

- 데이터 품질:
  데이터의 최신성, 정확성, 상호연계성 등을 확보하여 이를 사용자에게 유용한 가치를 줄 수 있는 수준
- 품질 관리 체계의 부재
  → 데이터 정확성 보장X
  → 데이터 자체의 신뢰도 저하
- 데이터 품질 체계 마련을 위한 실질적 노력 부족

데이터 불신뢰성 문제

- 데이터 불신뢰성 문제의 원인: 데이터 품질 관리 체계의 부재
- 데이터 불신뢰성 해결 방법: '데이터 라벨링' 절차 구축

데이터 라벨링과 신뢰도 문제

- 데이터 라벨링:
  데이터에 정확하고 의미 있는 레이블을 부여함으로써 모델의 학습 및 예측 능력을 향상시킴.
- 데이터 라벨의 불확실성 문제 존재

☑ 제조 데이터의 이슈: 데이터 라벨 부재

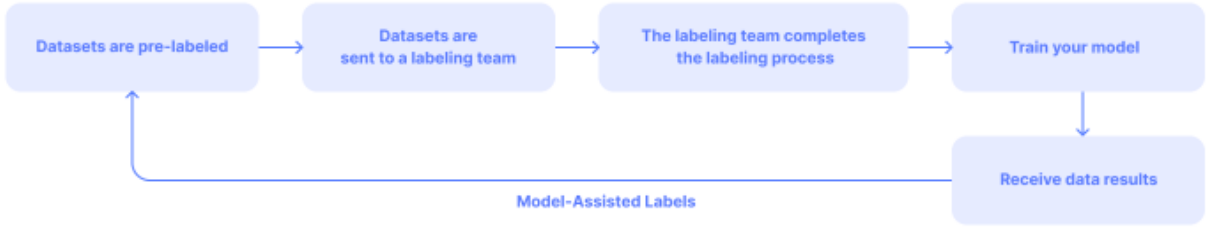
- 제조 공정에서 발생하는 데이터 분석의 목적: 전체 공정에서의 생산성/효율성 증진 및 이상 현상 모니터링을 통한 품질 향상
- 데이터 분석을 위한 필수 요소: **데이터 라벨**
    - 데이터 라벨 예시 – 양품/불량품
    - 분석용 제조 공정 데이터의 **라벨 – 부재**

- 데이터 라벨 부재 시 발생하는 문제
    - **모델 성능 평가 불가**: 정확도, 정밀도, 재현율과 같은 성능지표를 산출하는 데 필요한 라벨 정보가 없으므로 모델 성능 평가 불가
    - **새로운 데이터 수집 시 분류/예측력 저하 가능**: 모델이 학습한 내용을 기반으로 새로운 데이터에 대한 분류/예측이 어려움

- 데이터 라벨링을 위한 여러 방법론 대두
    - 엔지니어(사람) 의존적 방법론
    - 모델 의존적 방법론

☑ 기존 방법론 – Model Assisted Labeling

- 모델 의존적 방법론 – Model Assisted Labeling



- AI를 사전 학습시킨 후, 신규 입력되는 데이터를 AI모델이 자동으로 판단하여 라벨링 진행.
- 장점 – 사람의 자의적 개입이 이루어지지 않아 효율적인 라벨 부여 가능.

- 단점
  - 신규 입력되는 데이터의 특성이 기존 데이터와 상이할 때 라벨의 정확도 저하 가능.
  - 일부 사람의 개입이 필요한 상황에 개입 불가.
  - **사전 학습 할 데이터 라벨이 없는 상황에서는 활용 불가.**

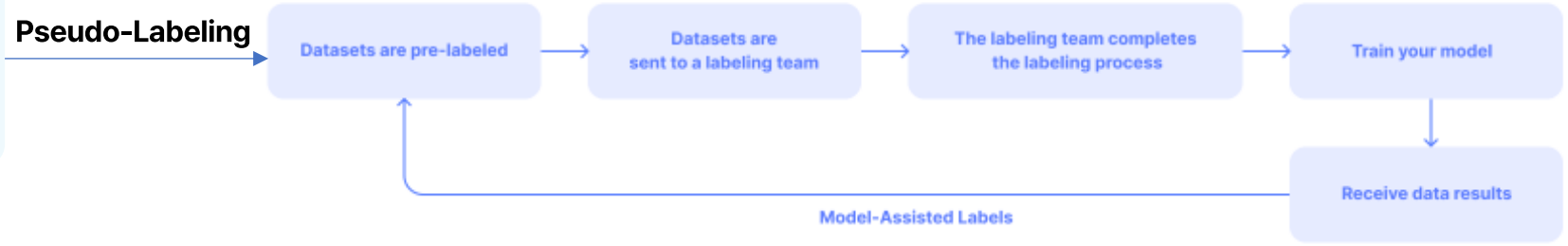
개선 방법론 – Hybrid Labeling Process

품질 데이터

- 양품 수, 불량품 수 , 총 수량

➔ **샘플 불량률** 추출

Pseudo-Labeling



Data-based Labeling

Model-based Labeling

샘플 불량률을 활용한 품질 예측 프레임워크

Data-based Labeling + Model-based Labeling

Hybrid Labeling Process

데이터 품질지수 측정

품질 데이터

- **데이터 크기**
  (136, 7)
- **양품 수, 불량품 수 , 총 수량**

공정 데이터

- **데이터 크기**
  (2939722, 21)
- **양품/불량품 라벨X**

학습 데이터

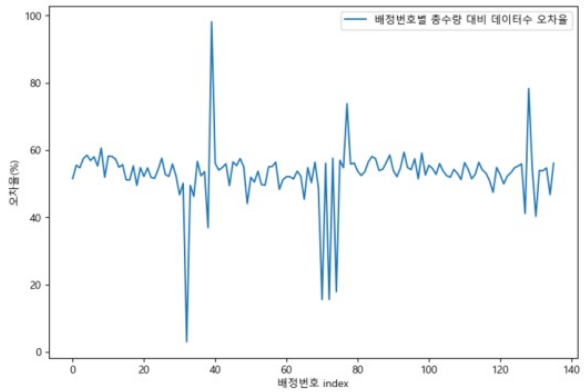
- **데이터 크기**
  (108, 38)

→

- Gartner의 품질지수

평가지표	검증 결과
일관성	O
완전성	O
유효성	O
정확성	X
유일성	X

- **정확성 및 유일성 부재**
  배정번호 별 품질 데이터의 총 수량 대비
  공정 데이터의 데이터 수 불일치 (오차 존재)



→

불량률 지표 기준값 도출

- **추정 불량률**
  품질 데이터로부터 도출된 불량률을 공정 데이터의 적용 시 실제 불량률이 아닌 **추정 불량률**로 활용
- **추정 불량률의 단계화**
  위험 단계 – 안정단계
  단계 구분기준치:
  위험 단계의 불량률 최솟값(0.000456)과 안정 단계의 불량률 최대값(0.000449)의 **중앙값(0.0004525)**

분석용 품질 데이터셋 구성

final_quality

배정번호	불량단계	데이터수	위험군개수
131033	0	17089	34.0
143256	0	20606	22.0
116862	0	21980	48.0
125637	0	14801	30.0
126519	1	13613	63.0
141893	0	41285	46.0
130868	0	19615	53.0
126569	0	26592	90.0

- 분석용 품질 데이터셋(파일명: final_quality.csv)
  - 배정번호 별 **위험군 개수**
  - 분석 모델에서의 종속 변수 중 하나로 활용

분석 데이터셋 및 변수 구성

- 활용 데이터셋
  - 전처리된 공정 데이터셋(preprocessed_data.csv)
  - 전처리된 품질 데이터셋(preprocessed_quality.csv)
- 변수
  - 독립변수: 배정번호 별 모든 feature
  - 종속변수: 배정번호 별 위험군 개수, 불량단계(0: 안정 단계, 1: 위험 단계)

분석 모델 개발

데이터셋 분리 및 가공

모델 아키텍쳐 설계

- 개별 데이터 분류기
- 배정번호 이상탐지 분류기

손실함수

- **군집 내부 거리 손실**
- **분류 불확실성 제어 손실**

학습 및 검증
하이퍼파라미터 설정

손실함수

군집 내부 거리 손실

- 불량률에서 비롯된 데이터의 품질이 '정상'에 치중되어 있으므로 불량률에 따른 입력 데이터 분포가 상이하다고 보기 어려움.
- 따라서, 해당 공정에서 전체적으로 공유하는 안정적인 공정의 입력 분포를 공유하고자 품질을 '정상'으로 분류한 데이터의 Latent Space 상에서의 위치를 활용.
- 유클리디안 거리 평균을 정상 데이터 응집력 손실로 설정.

$$\mathcal{L}_{cohesion} = \frac{\sum \left\| \boldsymbol{F}\left(\widehat{\mathbf{X}}'_{정상}\right) - \boldsymbol{E}\left[\boldsymbol{F}\left(\widehat{\mathbf{X}}'_{정상}\right)\right] \right\|_2}{N_{정상}}$$

분류 불확실성 제어 손실

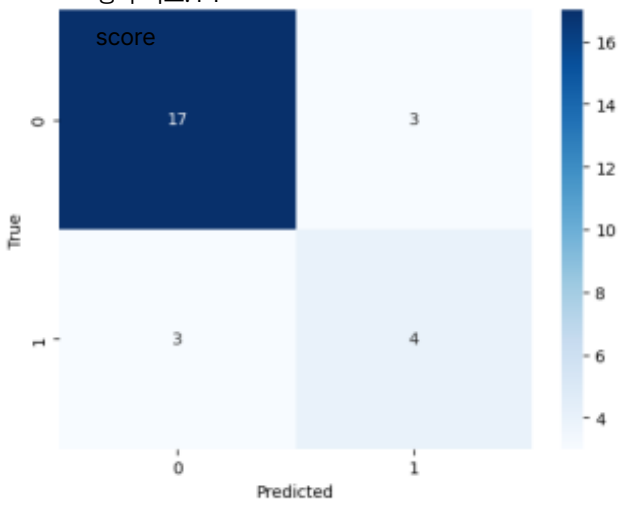
- 주어진 데이터의 라벨 중 신뢰도가 높은 것은 '불량 단계'이기에 '불량 단계' 예측 에러에 따라 손실의 가중치를 계단식으로 상이하게 설정함.
- 특히, 해당 데이터는 '위험'이 '안정' 대비 데이터의 수가 적지만 실제 작업 현장에서 발생하는 손실은 치명적이기에, '불량 단계 위험'을 '안정'으로 예측한 경우 손실 가중치를 최대로 설정.

$$\mathcal{L}_{proba} = \boldsymbol{I}(\boldsymbol{G_1}(\cdot) \geq \frac{N_{위험군}}{N}) \cdot (-\frac{\sum \log \boldsymbol{C}(\boldsymbol{F}(\widehat{\mathbf{X}}'_{정상}))}{\widehat{N}_{정상}} + \frac{\sum \log\left(1 - \boldsymbol{C}\left(\boldsymbol{F}\left(\widehat{\boldsymbol{X}}'_{위험군}\right)\right)\right)}{\widehat{N}_{위험군}}) \cdot \boldsymbol{W}_{anomaly} \cdot \boldsymbol{W}_{error}$$

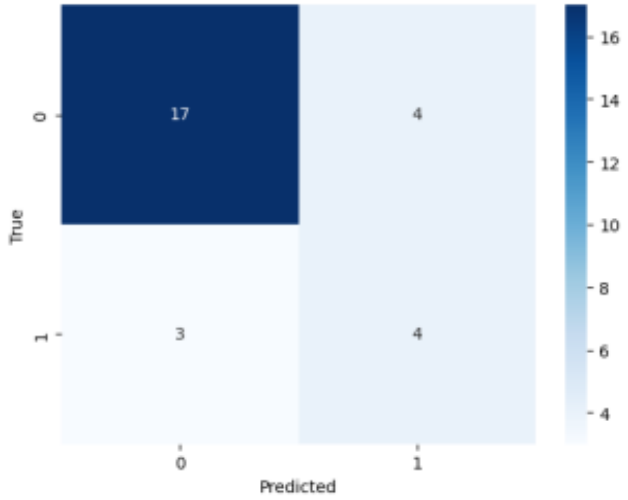
$$(\ \boldsymbol{W}_{anomaly} = \begin{cases} 1, & if\ \boldsymbol{G_2}(\cdot) = y \\ 10, & if\ y = 0, \boldsymbol{G_2}(\cdot) = 1 \\ 20, & if\ y = 1, \boldsymbol{G_2}(\cdot) = 0 \end{cases}, \boldsymbol{W}_{error} = \left| \boldsymbol{G_1}(\cdot) - \frac{N_{위험군}}{N} \right|, \mathbf{I}(\text{condition}) = \begin{cases} 1, & if\ \text{condition} \\ -1, & otherwise \end{cases})$$

분석 모델 학습 결과

- 평가 지표: F1 score



- Validation Confusion Matrix
  - **F1 Score: 0.7107**



- Test Confusion Matrix
  - **F1 Score: 0.6813**

- 개별 데이터에 대해 분류기의 추론 결과를 기반으로 위험군 비율 및 불량단계 추정 결과, 안정 및 위험 공정 분류 성능이 배정번호당 통계량을 입력값으로 사용한 '가이드라인'의 모델과 유사.

☑ 종합 결론 및 의의

- 제안 모델 학습 결과
  - 샘플 불량률을 활용하여 학습한 분류기가 개별 데이터의 위험군 여부 판단에 유의미한 결과를 보임.
  - 샘플 불량률을 활용한 제안 방법론이 **개별 데이터 라벨 추정 문제를 해결할 수 있음을** 시사.
- 의의
  - 품질 데이터로부터 산출된 추정 불량률을 활용한 데이터 라벨링 방법론 제안.
  - 사전 학습 할 데이터 라벨이 없는 상황에서는 활용 불가능했던 'Model Assist Labeling'방법론의 단점을 보완, 샘플 불량률을 활용한 사전 학습용 데이터 라벨 부여 모델 도출.
  - 데이터 라벨 부재한 상황에서 평가할 수 없었던 모델 성능을 산출 및 평가.

4. 결론

4-2. 기대효과

공정 효율성 및 생산성 향상

- 샘플 불량률만으로도 효과적으로 품질 예측이 가능하다는 점에서 데이터 수집 및 처리 비용을 절감 가능.
- 실시간 라벨 부여가 가능한 자동화 라벨링 프로세스를 활용 시 생산 품질 향상에 기여 가능.

라벨링 가이드라인 제공

- 중소 제조기업과 같이 데이터 전문 인력 부재로 인해 주관성과 오류가 늘어나는 상황에서 즉각적으로 활용 가능한 가이드라인으로 기능.
- 전문 인력 부재 상황에서도 라벨링 작업의 일관성과 품질 유지 가능.

범용성 높은 프레임워크

- 열처리 도메인에만 적용되는 모델이 아닌 다양한 산업 분야와 제조 공정에 대응되는 모델로, 높은 범용성을 갖는 프레임워크로 기능.

감사합니다