제4회 K-인공지능 제조데이터 분석 경진대회 보고서

프로젝트명	*작업자 친화적인 실시간 불량 예측 모델 개발*
팀명	*Family AI Kamp*
내용요약	**1. 프로젝트 개요** - 데이터 제공 기업은 일일 또는 주간 단위로 품질 이슈 현황을 파악하고 있으며 불량 원인을 수작업으로 분석하고 있어 즉각적인 대응이 가능하도록 제조 공정의 데이터를 분석하여 pass와 fail을 예측하는 프로그램을 제작한다. **2. 데이터 분석 및 전처리** - 총합 2,852,465개(행 92,015개, 열 31개)의 데이터로 구성되어 있다. - 품질 지수가 약 79.216%로, 분석이 불가능하여 데이터 특성을 파악하여 전처리를 진행하였다. - 전처리 결과 31개의 변수 중 21개를 선택하였고, 그 중 16개의 변수에서 치우침이 발생한는 것을 확인하였다. 비모수적 방법을 사용하는 것이 적절하다고 판단하여 XGBoost, LighGBM, CatBoost와 같은 트리기반 모델을 선정하였다. **3. 모델 학습 및 해석** - 각각의 베이스모델에 대하여 랜덤 서치를 진행하여 최적의 하이퍼파라미터를 선정하였고, 최종적으로 VotingClassifier에서 catboost : 1.0, lightGBM : 2.5, XGB : 1.0의 가중치로 학습을 진행하였다. - Pass, Fail의 클래스 불균형이 심하기 때문에 accuracy보다는 Fail 클래스에 대한 F1-score를 중점으로 판단하였다. - 최종적으로 TEST데이터셋에서 F1-score 96.5%의 예측률을 보여주었다. **4. 결론 및 시사점** - 제조 현장에서는 공정에 대한 지식은 있지만 AI에 대한 이해가 부족하기 때문에 AI 도입이 어려운 상황이다. 이를 해결하기 위해, 코드를 함수화하여 사용자가 코딩에 대한 지식이 없어도 변수만 변경하면 다양한 공정에 쉽게 적용할 수 있도록 프로그램을 개발하였다. - 불량 예측 및 공정 이슈를 해결하기 위해 투입되는 인력을 감소시켜 효율적인 인원 배분이 가능하다.

상기 본인(팀)은 위의 내용과 같이 제4회 K-인공지능 제조데이터 분석 경진대회 결과 보고서를 제출합니다.

2024년 11월 6일

팀장 : 최영민 (서명)

팀원 : 안태영 (서명)

팀원 : 허상호 (서명)

한국과학기술원장 귀중

□ 문제정의

ㅇ 공정(설비) 개요

▲ 공정 설비 예시 사진

- **다이캐스팅법**은 '다이(Die)'라고 불리는 금형에 금속 용탕을 고압으로 주입하여 정밀한 주물을 짧은 시간 내에 대량 생산한다.
- **Semi-solid 다이캐스팅법**은 용융 금속에 전단 응력을 가해 반응고 상태에서 주조하는 방식으로, 이 과정에서 전자 교반을 활용되므로 해당 공정은 Rheocasting 방식임을 알 수 있다.
- **다이캐스팅의 장점**은,
  1) 생산된 제품의 **완성도가 높아** 추가 가공이 거의 필요 없고,
  2) **비용이 저렴**하여 생산 속도가 빠르다는 장점이 있다.

ㅇ 공정(설비)상의 문제 현황 (Pain point)

- 경험에 의존한 설비 운용이 많아 **체계적인 공정 관리 시스템이 부족**하며, **데이터 기반의 예측 관리**가 잘 이루어지지 않는다.
- 데이터 제공기업의 경우 일일 또는 주간 단위로 품질 이슈 현황을 파악하고 있으며 **불량원인을 수작업으로 분석**하고 있다.
- 주조 공정은 노동집약적, 기술자산 비중이 높지 않은 특성을 보이며 특히 세계적인 추세와 다르게 우리나라는 선진국 대비 **낮은 인당 생산량** 등의 문제를 안고 있다.

○ 공정(설비)외 문제 현황 (Pain point)

- AI기술을 도입하기 위해서는 **초기 투자 비용**과 **전문 인력**의 필요로 많은 비용이 발생한다.
- 필요 데이터 수집을 위한 환경셋팅과 테스트 데이터 검증 후 실제 수집까지 시간이 요도될 수 있다.
- AI에 대한 이해 부족과 도입 후 기대할 수 있는 투자 수익률에 대한 불확실성은 AI 도입에 걸림돌이 된다.

○ 문제해결 장애요인

- 공정 데이터의 경우 회사의 기술력을 내포하고 있기 때문에 직접적인 외부 공개가 부족하다.
- 수집환경 셋팅과 분석, AI 적용까지 최소 수개월의 시간이 소요되며 인적, 물적 비용 발생이 예상된다.
- 작업자들은 AI에 대한 지식이 부족하고, AI 전문가는 공정에 대한 이해가 부족하다.

○ 극복방안

- 현장 전문가들과 지속적인 커뮤니케이션을 통해 공정에 대한 이해를 높이고 데이터분석을 통해 현장 전문가들이 미처 알지 못했던 공정의 특성에 대해 정보를 제공함으로써 상호 도움을 준다.
- 현재 공정 데이터는 수집되고 있지만, 품질 예측이나 수율 분석에 충분히 활용되지 못하고 있다. 따라서 수집된 데이터 분석과 머신러닝 알고리즘을 적용하여 불량 판정 및 공정 운영 최적화가 가능할 것으로 판단된다.

□ 제조데이터 정의 및 처리과정

ㅇ 제조 공정 및 데이터의 이해

- 주조 분야 : 다이캐스팅
- 수집 방법 : 주조 설비 내 PLC
- 수집 기간 : 2019년 01월 02일 ~ 2019년 03월 31일
- 데이터 개수 : 총합 2,852,465개(행 92,015개, 열 31개)

ㅇ 제조AI데이터셋 주요 변수 정의 (전처리 후)

속성(column)	설명	비고
count	일자별 제품 생산 번호	int64
working	가동여부	object
melten_temp	용탕온도	float64
production_CycleTime	제품생산 사이클 시간	int64
low_section_speed	저속구간속도	float64
high_section_speed	고속구간속도	float64
cast_pressure	주조압력	float64
biscuit_thickness	비스켓 두께	float64
upper_mold_temp1	상금형온도1	float64
upper_mold_temp2	상금형온도2	float64
lower_mold_temp1	하금형온도1	float64
lower_mold_temp2	하금형온도2	float64
lower_mold_temp3	하금형온도3	float64
sleeve_temperature	슬리브온도	float64
physical_strength	형체력	float64
Coolant_temperature	냉각수 온도	float64
EMS_operation_time	전자교반 가동시간	int64
tryshot_signal	사탕신호	object
mold_code	금형코드	object
heating_furnace	가열로	object
passorfail	양품불량판정	float64

[표 1] 분석에 사용된 21개의 변수 목록

ㅇ 주요 변수 기술 통계(전처리 후)

속성(column)	count	mean	std	min	50%	max	왜도
molten_temp	88569	718.7719	50.7567	0	728	735	-13.5018
production_cycletime	88569	122.2745	12.1565	0	121	462	5.5669
low_section_speed	88569	108.7370	7.9386	0	110	140	-7.4522
high_section_speed	88569	111.7759	9.2837	0	112	354	10.7329
cast_pressure	88569	324.9986	25.6672	139	330	345	-6.1023
biscuit_thickness	88569	50.5012	15.4915	0	50	422	22.0015
upper_mold_temp1	88569	186.9405	45.7599	19	195	337	-0.5306
upper_mold_temp2	88569	166.6410	28.3461	15	173	243	-0.5620
lower_mold_temp1	88569	205.3395	53.2084	20	211	367	-0.2692
lower_mold_temp2	88569	200.9323	43.5820	26	199	499	0.0845
lower_mold_temp3	88569	1441.9420	54.6192	638	1449	1449	-8.0316
sleeve_temperature	88569	417.0282	115.5642	33	458	1449	-0.7093
physical_strength	88569	699.7702	48.1839	0	703	736	-13.3081
Coolant_temperature	88569	32.1588	2.6030	17	32	48	-0.0328
EMS_operation_time	88569	18.3018	8.4260	0	23	25	-1.2893

[표 2] 분석에 사용된 16개의 변수 기술 통계량

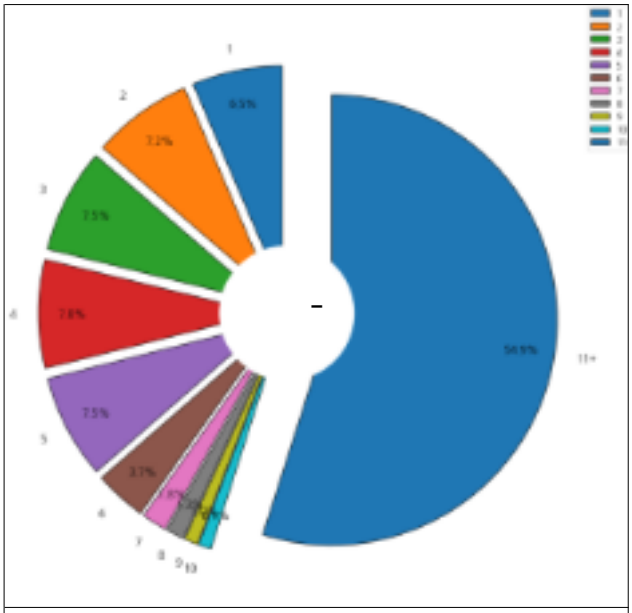
속성(column)	count	mean	std	min	50%	max	왜도
count	88569	148.9747	87.7641	1	147	334	0.0828
working	88569	0.9997	0.0181	0	1	1	-55.2377
tryshot_signal	88569	0.0190	0.1365	0	0	1	7.0483
mold_code	88569	3.6324	2.4405	0	5	6	-0.4680
heating_furnace	88569	1.2823	0.8256	0	2	2	-0.5640
passorfail	88569	0.0397	0.1952	0	0	1	4.7159

[표 3] 분석에 사용된 5개의 변수 기술 통계량

- [표 2], [표 3]은 분석을 위해 선정된 변수들의 기술 통계량을 나타내고 있다.
- [표 2]에서 대부분의 변수들이 치우침을 보이는 것을 확인할 수 있다.
- 대부분의 변수들이 치우침을 보이며, 데이터의 특성을 고려할 때 비모수적 방법인 '트리 기반 모델'을 사용하는 것이 적절하다고 판단된다. * | 왜도 | < 0.5 일 때 정규분포와 근사한 형태를 보인다고 할 수 있다.

○ 제조 데이터를 기반으로 공정 특성 유추

- '전자교반 3라인 2호기'는 동시에 2개의 mold로 작업이 가능하며, 각각의 작업 수량(count)이 기록된다.
- 공정은 연속적으로 진행되며, 통상적으로 하루 2회(07, 19시) 운전이 멈춘다
- 장비가 재가동될 때 새롭게 count가 진행되며, 장비 재가동 후 안정화까지 보통 5~6회의 작업이 소요되는 것으로 확인된다.



▲ [그림 2] count 별 fail 비율 파이차트

○ 결측치 제거

- passorfail에서 발생한 1개의 결측치는 제거하였다.
- heating_furnace와 molten_volume이 모두 없는 경우는 적은 비율(2.46%)로 인하여 제거하였다.
- tryshot_signal은 장비가 준비되지 않았을 때 발생하는 신호로, 정상 상태에서는 발생하지 않으므로 'No'로 채워주었다.
- lower_mold_temp3, upper_mold_temp3, molten_temp의 결측치 비율은 최대 2.5%로, mold_code별 최빈값으로 대체하였다.
- heating_furnace와 molten_volume은 각각 약 50%의 결측치를 가지고 있지만 서로 대응되는 형상을 보여준다. 데이터의 특성을 고려할 때, molten_volume은 heating_furnace의 A, B가 아닌 다른 방법(Others)으로 간주할 수 있으므로, heating_furnace의 결측치에 'C'를 넣어주고 molten_volume은 삭제하였다.

ㅇ 유의하지 않은 데이터 제거

- 고유값을 가지는 line, name, mold_name, emergency_stop 컬럼을 제거한다.
- time, date, registration_time은 데이터 분석에 유의미하지 않다고 판단하여 제거한다.

ㅇ 문자형 데이터 인코딩

- OrdinalEncoder를 통해 Object 타입의 데이터를 숫자로 인코딩하였다.

ㅇ 이상치 제거

- 해당 데이터는 PLC를 통해 공정 데이터를 수집하고 있으며, 순간적인 변화(Hunting)가 기록될 수 있다.
- pass 데이터와 fail 데이터가 각각 정상 범위와 이상치 범위로 분포할 것이라 예상하여, 공정에서 발생할 수 없는 극단적인 값들만 제거하기 위해 LocalOutlierFactor (LOF) 방식을 이용하여 이상치를 제거하였다.
- LOF 방식은 데이터의 밀집도를 측정해 지역적 이상치를 찾아내는 비지도 학습 기법으로, 데이터 간 거리를 기반으로 이상치를 찾아낸다. 공정상의 극단적 이상치 제거를 위해 n_neighbors=10으로 설정하였다.

ㅇ 양품집단과 불량집단간의 T-test

- t-검정을 통해 각 변수의 pass와 fail 그룹 간 평균 차이를 비교하였다. 각 그룹의 평균은 해당 그룹에 속하는 데이터의 수치적 값을 기반으로 계산되었다.
- 분석 결과, p-value가 0.05보다 작은 경우, 두 그룹 간에 유의미한 차이가 있다고 판단하며, 이때 특정 피처가 공정 결과에 영향을 미친다고 해석할 수 있다.
- 검정 결과 '설비작동사이클시간', '상금형온도3'은 제거하였다.

ㅇ 학습-평가 데이터 분리(Train-Test-Split)

- 머신러닝 모델을 학습 및 평가하기 위해 훈련 세트(train set)와 테스트 세트(test set)를 8:2비율로 나눠 진행하였다.
- 두 집단 모두에서 pass와 fail의 비율이 균등하게 유지되도록 조정하여 학습을 진행하였다.

ㅇ 데이터 품질 지수(Data Quality Index)

- 데이터셋의 품질을 평가하기 위해 완전성, 유일성, 유효성, 일관성, 무결성을 종합적으로 평가하여 품질 지수를 산정하였다.
- 공정데이터 특성을 고려하여 무결성과 완전성에 높은 가중치를 부여하였

데이터 품질 지수	설명
완전성 품질 지수	각 칼럼에 누락값이 없어야 한다. 데이터 누락이 생산성 품질에 영향을 주므로 큰 가중치를 부여한다. ((1-결측치)/전체 데이터 수) * 100
유일성 품질 지수	데이터 항목은 유일해야하며 중복되어서는 안된다. 재고 관리나 생산 계획에 혼란을 초래하며 비교적 낮은 가중치를 부여한다. ((유일한 데이터 수)/전체 데이터 수) * 100
유효성 품질 지수	데이터 항목은 정해진 데이터 유효범위 및 도메인을 충족해야 한다. 데이터의 실용성을 보장하기 위해 중간 정도의 가중치를 부여한다. (유효성 만족 데이터 수/전체 데이터 수)*100
일관성 품질 지수	데이터의 표현되는 형태가 일관되게 정의되고 서로 일치해야한다. 공정 간 통합을 위해 중요하며, 중간에서 높은 가중치를 부여한다. (일관성 만족 데이터 수/전체 데이터 수)*100
무결성 품질 지수	데이터의 관계가 손상되지 않고 올바르게 연결되어 있어야 한다. 전체 시스템의 신뢰성을 위해 매우 중요하므로, 높은 가중치를 부여한다. (1-(유일성, 유효성, 일관성 지수중 100%가 아닌 지수 개수 / 3))*100
가중치 지수	각 지표의 중요도를 부여하고, 이를 바탕으로 종합 지수를 계산한다. 무결성 > 완전성 > 일관성 > 유효성 > 유일성 순으로 가중치를 부여한다. 품질 지수 * 가중치
데이터 품질 지수	데이터가 얼마나 신뢰할 수 있고 분석 및 의사결정에 적합한지 평가하는 지표 Σ 가중치 지수

[표 4] 데이터 품질 지수의 지표별 정의 및 산출 방식

- 전처리 전 데이터의 가중치 지수는 약 79.217%로 '보통' 등급으로 평가되었다. 이는 데이터의 품질이 어느 정도 만족스럽지만, 분석을 진행하기 위해서는 데이터 정제가 필요하다는 것을 나타낸다.

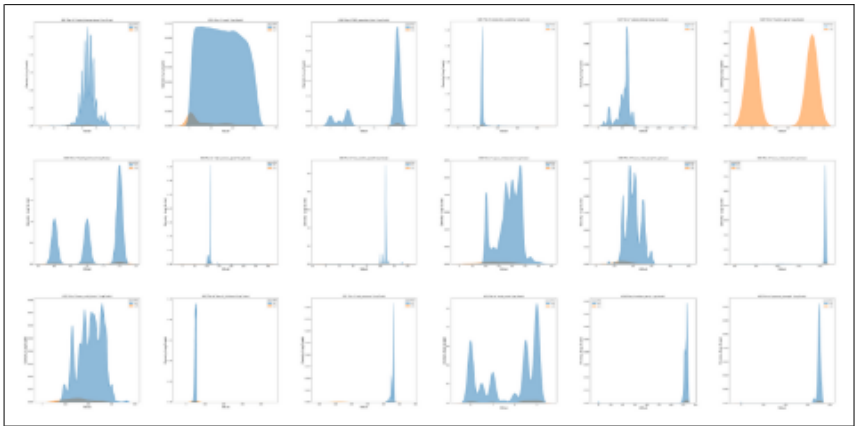
구분	품질지수	가중치	가중치 지수	오류율
완전성(누락)	97.51%	25.00%	24.377%	0.623%
유일성(중복)	99.97%	10.00%	9.997%	0.003%
유효성(유효)	98.95%	15.00%	14.843%	0.158%
일관성(표현)	100%	20.00%	20%	0%
무결성	33.33%	30.00%	9.999%	20.001%
품질지수	85.95%	100.00%	79.216%	14.05%(20.784%)

[표 5] 전처리 전 데이터의 품질 지수 및 오류율

- 전처리 후 데이터의 가중치 지수는 100.00%로 '매우 좋음'등급으로 평가되었다. 이는 데이터가 완벽하게 정제되어 분석을 진행하기에 적합한 상태가 되었음을 확인할 수 있다.

구분	품질지수	가중치	가중치 지수	오류율
완전성(누락)	100.00%	25.00%	25.00%	0.00%
유일성(중복)	100.00%	10.00%	10.00%	0.00%
유효성(유효)	100.00%	15.00%	15.00%	0.00%
일관성(표현)	100.00%	20.00%	20.00%	0.00%
무결성	100.00%	30.00%	30.00%	0.00%
품질지수	100.00%	100.00%	100.00%	0.00%

[표 6] 전처리 후 데이터의 품질 지수 및 오류율



▲ [그림 3] 전처리 후 데이터의 분포

□ 분석모델 개발

ㅇ 제조 공정 분석을 위한 최적의 모델 개발

- 데이터 제공 기업은 품질 이슈를 일일 또는 주간 단위로 파악하여 즉각적인 대응이 어려운 상황이다.
- 공정 과정에서 불량 제품 발생 시 작업자가 쉽게 인지할 수 있도록 AI 모델 기반의 공정 파라미터 분석에 집중하였다.
- 공정 데이터의 특성을 반영한 데이터 전처리를 바탕으로 XGBoost, LightGBM, CatBoost를 베이스 모델로 하는 VotingClassifier를 활용하여 앙상블 기법을 설계하였다.

ㅇ XGBoost(Extreme Gradient Boosting) 소개

XGBoost는 병렬 처리를 지원하여 대규모 데이터셋에서도 빠른 학습이 가능하며, 이는 대량의 데이터 처리에 매우 적합한 특징이다. 이 모델은 L1 및 L2 정규화 기법을 내장하고 있어 과적합을 방지하고 모델의 복잡도를 줄이는 데 효과적이다. 또한, XGBoost는 부스팅 방식으로 샘플 데이터를 추출하여 여러 개의 분류기를 생성하되, 이전 분류기의 학습 결과를 바탕으로 다음 분류기의 학습 데이터 가중치를 조절하여 학습을 진행한다. 이러한 특성을 통해 점진적으로 오차를 줄일 수 있으며, 결과적으로 예측 성능을 극대화할 수 있다.

ㅇ LightGBM(Light Gradient Boosting Machine) 소개

LightGBM은 손실이 큰 리프부터 선택적으로 확장하는 리프 중심의 트리 성장 방식을 채택하여, 노드를 분할할 때 전체 데이터에서 가장 최적의 분할을 찾아 리프를 성장시킨다. 이러한 접근 방식은 일반적인 균형 분할 방식보다 더 빠르게 수렴하면서도 낮은 손실을 유지하는 장점을 제공한다. 또한, Gradient-based One-Side Sampling (GOSS) 기법을 활용하여 학습 과정에서 중요하지 않은 데이터 포인트를 줄이고, 중요한 데이터에 더 많은 가중치를 부여함으로써 학습의 효율성을 높인다.

○ CatBoost(Category Boosting) 소개

CatBoost는 범주형 데이터와 불균형 데이터 처리에서 뛰어난 성능을 발휘한다. 이 모델은 대칭 트리를 활용하여 모든 노드가 동일한 조건으로 분할되므로, CPU 효율성을 극대화하고 예측 시간을 단축하며 과적합을 방지하는 데 도움을 준다. 또한, CatBoost는 순서 부스팅(Ordered Boosting) 방식을 채택하여 예측 시프트 문제를 해결한다. 이를 위해 데이터의 순서를 무작위로 섞고 일부 데이터로 모델을 학습한 후, 나머지 데이터로 잔차를 계산하여 과적합을 줄인다.

○ VotingClassifier 소개

- VotingClassifier는 다양한 모델을 결합하여 예측 성능을 향상시키는 앙상블 기법으로, 하드 보팅과 소프트 보팅의 두 가지 방법이 있다. 소프트 보팅은 각 모델의 확률 값을 평균내어 가장 높은 확률 값을 가진 클래스를 선택하는 방식으로, 더 세밀한 예측을 가능하게 한다. 이렇게 함으로써 모델 다양성을 확보하고 서로 다른 모델의 장점을 결합하여 성능과 일반화 성능을 높일 수 있다.
- 또한, 특정 모델이 더 중요한 경우, 해당 모델에 더 높은 가중치를 부여하여 데이터를 더 잘 분류하는 모델을 만들 수 있다. 이를 통해 전체 모델의 성능을 향상시키고, 특정 문제에 대한 모델의 기여도를 조정하는 유연성을 가진다.

○ 해당 AI방법론(알고리즘) 선정 이유

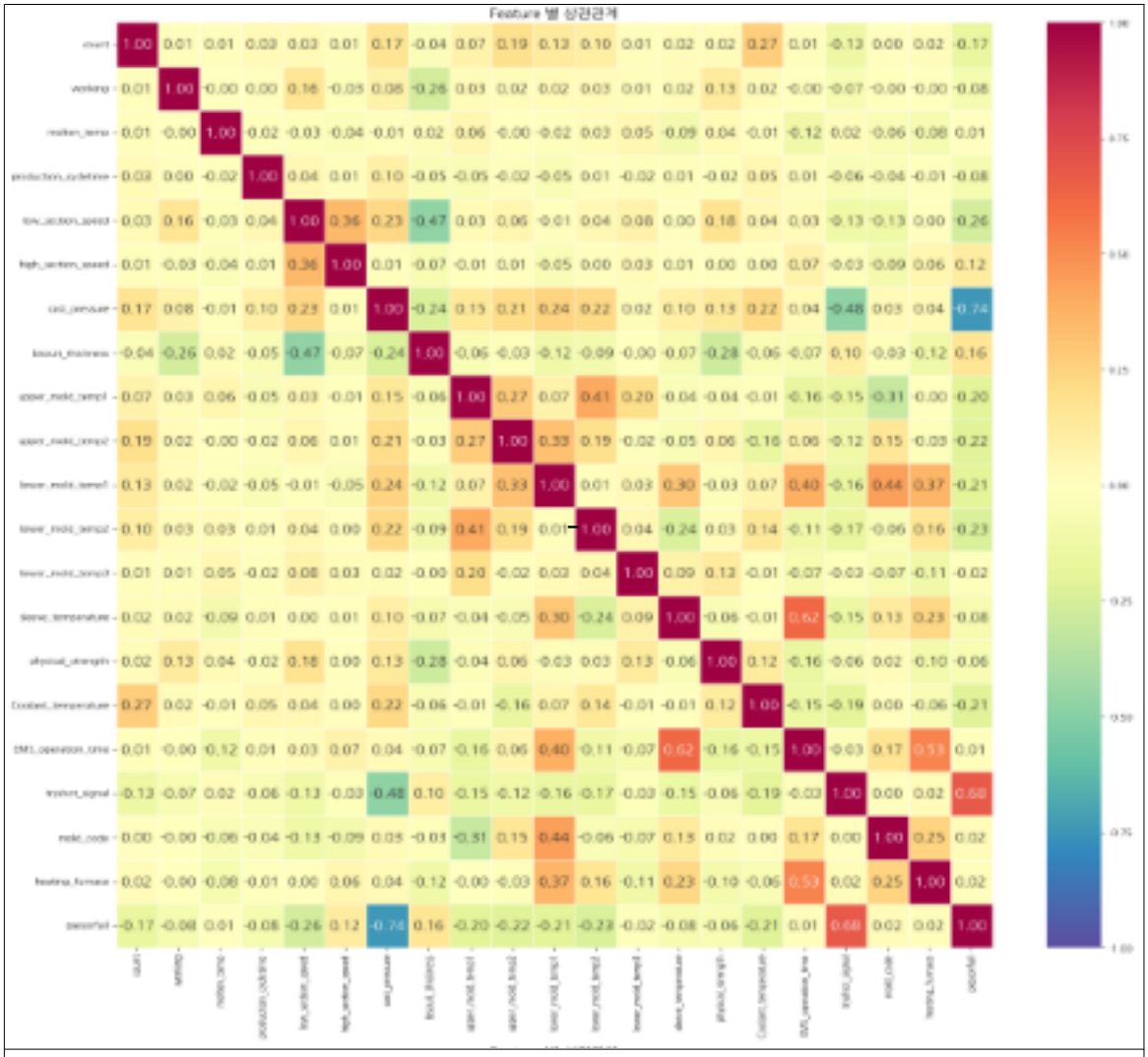
- 선택 AI방법론 : VotingClassifier(XGBoost, LightGBM, CatBoost)
- AI 데이터셋 전처리 결과, 21개의 변수 중 16개 변수에서 치우침이 발생하였고, 그 중 11개 변수는 매우 강한 치우침을 보였다. 이러한 데이터 특성을 고려할 때, 비모수적 방법을 사용하는 것이 적절하다고 판단하였다. 따라서 XGBoost, LightGBM, CatBoost와 같은 트리 기반 모델을 선정하였고, 이 모델들의 예측 성능을 극대화하기 위해 VotingClassifier를 이용하여 최종 분석을 진행하였다.

- XGBoost: 이전 분류기의 학습 결과를 바탕으로 다음 분류기의 학습 데이터 가중치를 조절하여 학습을 진행한다. 이러한 특성을 통해 점진적으로 오차를 줄일 수 있으며, 결과적으로 예측 성능을 극대화할 수 있다.
- LightGBM: 리프 중심의 트리 성장 방식과 GOSS(Gradient-based One-Side Sampling) 기법을 활용하여 학습의 효율성을 높인다. 이로 인해 대량의 데이터를 빠르게 처리할 수 있다.
- CatBoost: 범주형 데이터와 불균형 데이터 처리에서 뛰어난 성능을 발휘한다. Pass와 Fail의 클래스 불균형이 심하게 발생하고 있기 때문에 이 모델을 선정하였다.
- VotingClassifier: 각 베이스 모델이 서로 다른 학습 방법론과 특성을 가지고 있어 모델의 다양성을 확보할 수 있다. 또한, 서로 다른 가중치를 할당함으로써 보다 세밀한 예측이 가능하며, 앙상블 기법을 통해 과적합을 방지하는 데 도움이 된다.
- 이러한 이유로 XGBoost, LightGBM, CatBoost를 베이스 모델로 활용하고, VotingClassifier를 적용하여 최종 분석을 수행함으로써 치우침이 있는 데이터를 보다 효과적으로 분석하고 예측할 수 있을 것으로 예상된다.

□ 분석결과 및 시사점

ㅇ 전처리 데이터셋을 활용한 학습 진행 과정

- [그림]과 같이 전처리가 끝난 21개의 변수에 대하여 상관관계를 확인하였다.
- 상관계수는 -1과 1사이의 값을 가지며, 1에 가까울수록 강한 양의 상관관계를, -1에 가까울수록 강한 음의 상관관계를 나타낸다.



▲ [그림 4] 21개 변수간의 상관관계를 나타낸 히트맵

- 종속변수 [passorfail]은 독립변수 [cast_pressure]와 -0.74의 강한 음의 상관관계가 나타나며, 독립변수 [tryshot_signal]과 0.68의 중간 정도의 양의 상관관계가 나타나고 있다.
- 종속변수 [EMS_operation_time]는 종속변수 [sleeve_temperature]와 0.62, 종속변수 [heat_furance]와 0.53의 중간 정도의 양의 상관관계가 나타나고 있다.

ㅇ 전처리 데이터셋을 활용한 학습 결과 및 성능 확인

- 훈련 세트와 테스트 세트를 8:2비율로 나누고, 두 집단 모두에서 pass와 fail의 비율이 균등하게 유지되도록 조정하여 학습을 진행하였다.
- 베이스 모델은 랜덤 서치를 통해 하이퍼파라미터를 선정하였으며, 최적의 파라미터는 다음과 같다.

	n_estimators	max_depth	Learning_rate	Boosting_type	l2_leaf_reg	iterations	depth
lightGBM	400	-1	0.1	gbdt	-	-	-
XGBoost	300	12	0.1	-	-	-	-
catBoost	-	-	0.1	-	5	1000	8

[표 7] 랜덤 서치를 통해 얻은 최적의 하이퍼파라미터

- 각각의 베이스 모델에 최적의 하이퍼미터를 적용 후 최종 분석 모델인 VotingClassifier에서 예측을 진행한 결과는 [표]와 같다. 이때, 모델의 가중치는 catboost : 1.0, lgbm : 2.5, xgb : 1.0으로 선정하였다.

classification report :

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	17011
1.0	0.98	0.95	0.96	703
accuracy			1.00	17714
macro avg	0.99	0.97	0.98	17714
weighted avg	1.00	1.00	1.00	17714

classification report :

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	17011
1.0	0.97	0.93	0.95	703
accuracy			1.00	17714
macro avg	0.99	0.97	0.98	17714
weighted avg	1.00	1.00	1.00	17714

classification report :

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	17011
1.0	0.97	0.94	0.95	703
accuracy			1.00	17714
macro avg	0.98	0.97	0.98	17714
weighted avg	1.00	1.00	1.00	17714

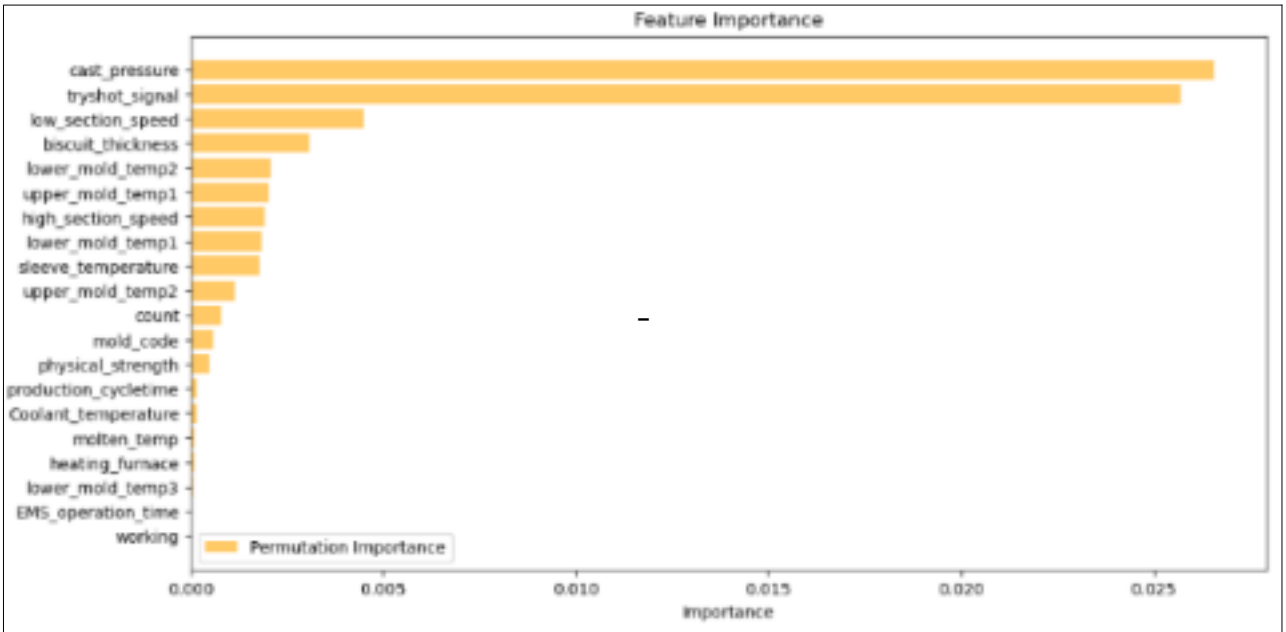
classification report :

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	17011
1.0	0.98	0.95	0.97	703
accuracy			1.00	17714
macro avg	0.99	0.97	0.98	17714
weighted avg	1.00	1.00	1.00	17714

▲ [그림 4] 모델별 classification report

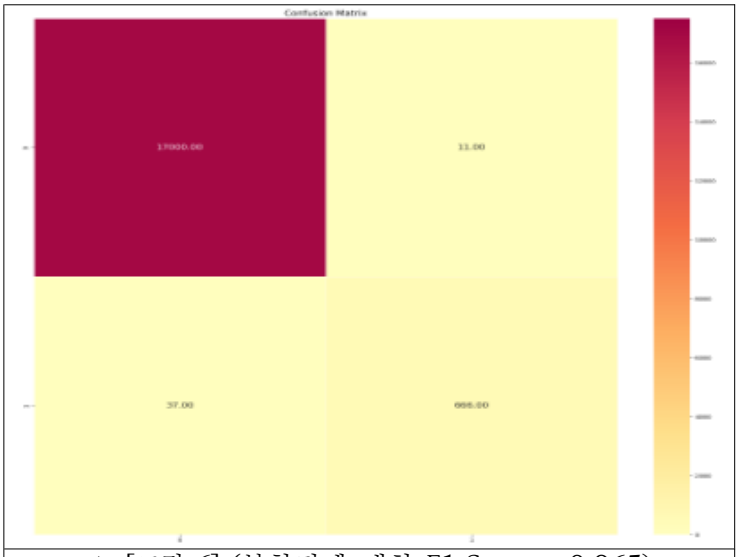
(좌 상단) Lgbm, (좌 하단) XGB, (우 상단) catBoost, (우 하단) VotingClassifier

- 최종 모델 Confusion Matrix를 확인한 결과, 37개의 fail 데이터를 잘 못 분류한 것을 확인할 수있다. pass와 fail 클래스가 불균형한 분포를 가지고 있기 때문에, Accuracy는 성능 평가 지표로서 큰 의미가 없으며, fail 클래스에 대한 F1-score를 확인한 결과 약 0.965의 성능을 보여주고 있다.



▲ [그림 5] (Model의 Permutation Importance)

- 최종 모델의 Feature Importance를 확인한 결과, 주조 공정에 많은 영향을 미칠 것으로 예상했던 온도와 압력 변수들이 높은 비중을 차지하고 있다.



▲ [그림 6] (불합격에 대한 F1-Score : 0.965)

ㅇ AI데이터셋 분석 시사점

- 최종 모델의 분석 결과, 불량 예측에 많은 영향을 미치는 주요 인자들은 Permutation Importance 기법을 통해 확인할 수 있다. 이 분석을 통해 공정에서 관리 가능한 변수로는 cast_pressure, low_section_speed, biscuit_thickness, upper_mold_temp1 등이 중요한 영향을 미친다는 것을 확인할 수 있었다. 또한, 공정 또는 장비 문제를 의미하는 tryshot_signal 변수가 불량 예측에 상당한 영향을 주는 중요한 요소로 나타났다.
- 이러한 분석 결과를 바탕으로, 공정에서는 cast_pressure 변수에 대한 세밀한 통제가 필요하며, 이를 위한 철저한 관리 방법을 도입이 필요하다
- 제조 공정의 다른 장비에 AI 분석을 적용하면, 변수별 수준에 따른 장비의 차이를 확인할 수 있다. 이러한 방법은 눈에 보이지 않는 변화를 빠르게 감지할 수 있어, 문제를 해결을 위한 인력 투입을 효율적으로 절감할 수 있을 것으로 예상된다.
- Permutation Importance 기법을 통해 확인한 결과, count 변수는 불량 예측에 중요한 영향을 미치지 않는 것으로 나타났다. 그러나 데이터를 분석한 결과, 장비 재가동 후 안정화까지 5~6회의 작업이 필요한 것을 확인하였다. 이를 바탕으로, 장비의 정상화를 위한 Dummy 작업을 진행한다면 비용 절감과 함께 total loss를 줄일 수 있을 것으로 예상된다.
- 장비상의 PLC에서 기록되는 데이터를 이용하여 분석을 진행하였다. 향후 각종 불량 유형을 이후 공정에서 라벨링하게 된다면, 이를 통해 불량 원인을 보다 정확하게 판단할 수 있을 것이라 예상된다.

□ 중소제조기업에 미치는 파급효과

ㅇ 손쉬운 사용을 통해 다양한 공정에 적용

- 제조 현장에서는 공정에 대한 지식은 있지만 AI에 대한 이해가 부족해 AI 도입이 어려운 상황이다. 이를 해결하기 위해, 본 팀은 코드를 함수화하여 사용자가 코딩 지식 없이도 변수만 변경하면 다양한 공정에 쉽게 적용할 수 있도록 프로그램을 개발했다.

ㅇ 업무 환경 개선

- 품질관리를 사람이 직접 하는 것이 아닌 공정 데이터를 통해 예측함으로서 품질 관리 및 유지보수가 개선되어 전반적인 생산성 향상을 기대할 수 있을 것이다.
- 중소 기업의 제조 경쟁력을 강화하여 지속적인 성장을 이끌어 낼 수 있을 것이다.

ㅇ 생산성 증가 및 품질 관리 비용 절감

- 공정 최적화를 통해 비용을 절감하고 생산 및 출시 기간을 단축할 수 있을 것이다.
- 기계 센서 데이터를 바탕으로 AI 기반 품질 검사를 통해 불량률을 최소화하고 검사 시간을 단축시킬 수 있을 것이다.
- 추가적으로 설비의 고장 패턴을 분석한다면 고장을 사전에 감지하고 예방할 수 있어 설비 유지보수 비용을 절감할 수 있을 것이다.
- AI 분석을 통해 불량 공정의 원인을 경험이 아닌 데이터를 바탕으로 찾을 수 있게 될 것이다. 즉, 시행착오를 줄이고 공정의 효율성 증가 및 제품 개발 비용을 절감에 효과적일 것으로 예상된다.