LS BIGDATA SCHOOL

4주차 프로젝트

2조 결과 분석 보고서

정하연 강하연 김지유 박민서 이채은 정재성

목차

프로젝트 주제 선정

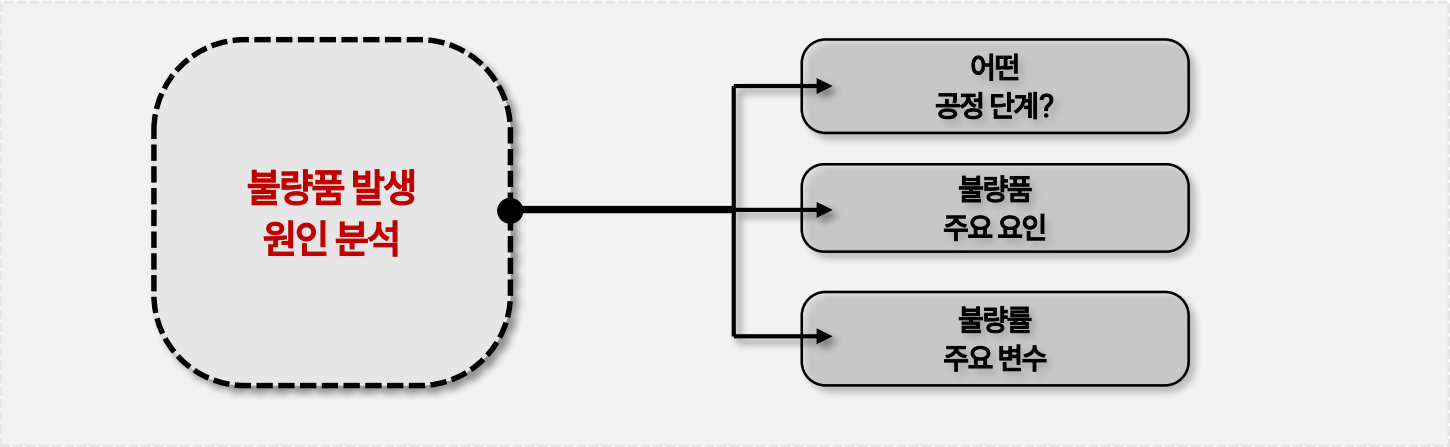
프로젝트 주제

선정 배경

제조업 주조 공정 → 공정의 효율성 및 품질관리 → 기업의 경쟁력 → 불량품 및 불량률 → 최소화 필요 → AI 데이터셋 활용 불량품 요인 분석 → 해결 방안 탐색

프로젝트 주제 선정

문제 정의



분석구조도

탐색적 데이터 분석 시도 → 피처 엔지니어링 → 예측모델 생성

확보한 데이터를 통한
결측치 보간

각 데이터간의 상관성 분석

프로젝트 수행 절차 및 팀 구성

프로젝트 수행 절차

6월 3일 – 6월 5일

프로젝트 주제 선정 및 데이터 분석
- 프로젝트 주제 선정
- 데이터 현황 파악

탐색적 분석 (EDA)
- 데이터 시각화 (주요 변수 파악)
- 불량 요인 탐색

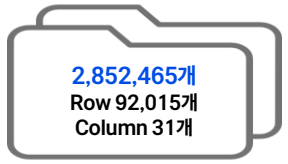
예측 모델 개발 및 서비스 제공
- 데이터 전처리
- 예측 모델 개발
- 향후 발전 방향 제시

팀 구성

훈련생	주 역할	담당 업무
정하연 (팀장)	EDA, 데이터 시각화	**EDA를 통한 불량요인 파악 및 발표**
강하연	데이터 전처리, 모델링	**최적의 모델 탐색을 통한 성능 개선 탐구**
김지유	EDA, 데이터 시각화	**데이터 시각화를 통한 데이터 탐구**
박민서	EDA, PPT 제작	**주제 선정 및 PPT 제작**
이채은	EDA, 데이터 시각화	**모델 평가 및 결론 도출**
정재성	데이터 전처리, 모델링	**결측치 처리, 최적의 모델 탐색 및 발표**
김웅기, 박종률	멘토	**질의응답 및 방향성 제시**

데이터 현황 파악

데이터 명세서



속성(column)	설명
line	작업라인
name	제품명
mold_name	금형명
time	수집시간
date	수집일시
count	**일자별 제품 생산 번호**
working	가동여부
emergency_stop	비상정지
molten_temp	용탕온도
facility_operation_CycleTime	설비 작동 사이클 시간
proudction_CycleTime	제품생산 사이클 시간
low_section_speed	저속구간속도
high_section_speed	고속구간속도
molten_volume	용탕량
cast_pressure	주조압력
biscuit_thickness	비스킷 두께
upper_mold_temp1	상금형온도1
upper_mold_temp2	상금형온도2
upper_mold_temp3	상금형온도3
lower_mold_temp1	하금형온도1
lower_mold_temp2	하금형온도2
lower_mold_temp3	하금형온도3

속성(column)	설명
sleeve_temperature	슬리브 온도
physical_strength	형체력
coolant_temperature	냉각수 온도
EMS_operation_time	전자교반 가동시간
registration_time	등록일시
passorfail	양품불량판정
tryshot_signal	사탕신호
mold_code	**금형코드**
heating_furnace	가열로

싸이클 분리
'trial'
파생 변수 생성

mold_code	count	* trial
8722	1	3
8722	2	3
8722	···	3
8722	288	3
8722	289	3
8722	290	3
8722	1	4
8722	2	4
8722	3	4

데이터 현황 파악

속성(column)	설명
line	작업라인
name	제품명
mold_name	금형명
time	**수집시간**
date	수집일시
count	일자별 제품 생산 번호
working	가동여부
emergency_stop	비상정지
molten_temp	용탕온도
facility_operation_CycleTime	설비 작동 사이클 시간
proudction_CycleTime	제품생산 사이클 시간
low_section_speed	저속구간속도
high_section_speed	고속구간속도
molten_volume	용탕량
cast_pressure	주조압력
biscuit_thickness	비스킷 두께
upper_mold_temp1	상금형온도1
upper_mold_temp2	상금형온도2
upper_mold_temp3	상금형온도3
lower_mold_temp1	**하금형온도1**
lower_mold_temp2	**하금형온도2**
lower_mold_temp3	하금형온도3

속성(column)	설명
sleeve_temperature	슬리브 온도
physical_strength	형체력
coolant_temperature	냉각수 온도
EMS_operation_time	전자교반 가동시간
registration_time	등록일시
passorfail	양품불량판정
tryshot_signal	사탕신호
mold_code	금형코드
heating_furnace	가열로

파생 변수

gap	약 2분 간격이 아닌 3~4분의 이상일 경우, 온도가 떨어지는 경향 발견 시간간격을 기록하는 'gap'열 추가
gap_sign	3~4분 이상일 경우, 시간 차이가 170초 이상 시 1, 아니면 0을 기록

count 변수를 이용한 *trial 파생 변수 생성

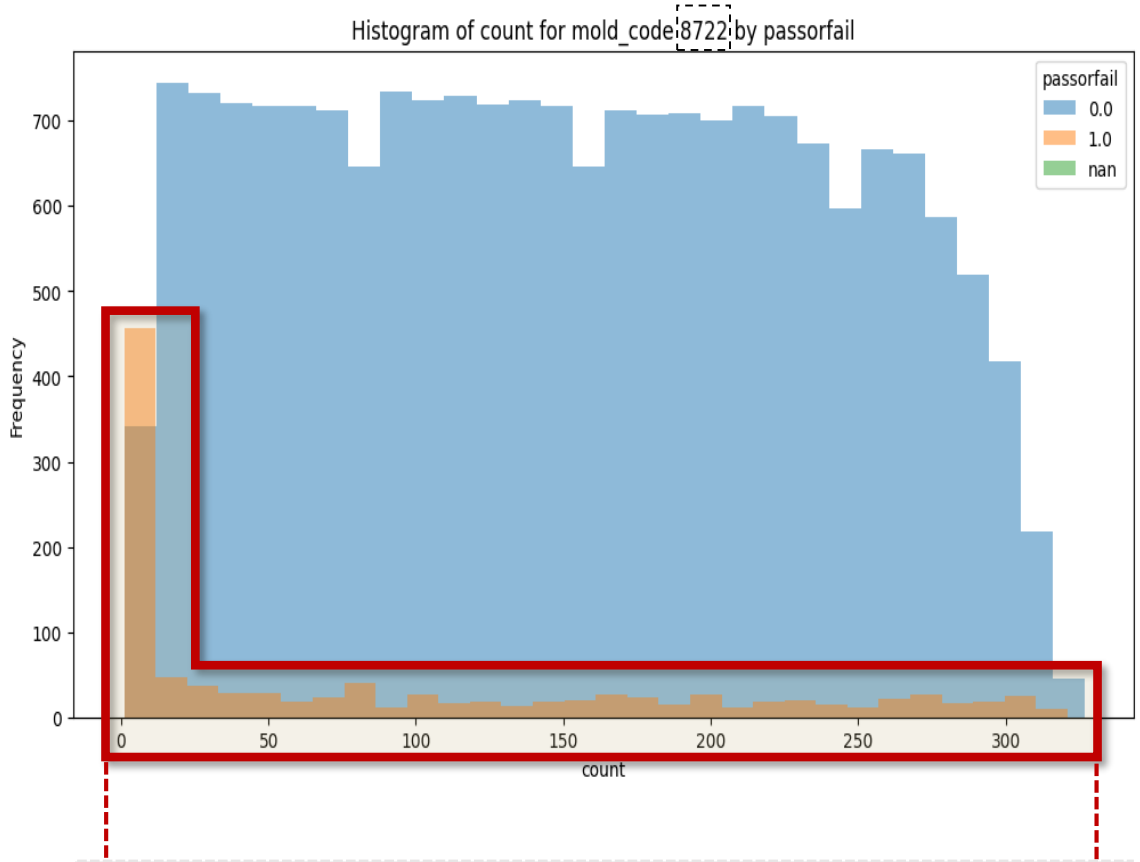
mold_code	count	trial
8722	1	3
8722	2	3
8722	···	3
8722	288	3
8722	289	3
8722	290	3
8722	1	4
8722	2	4
8722	3	4

① **'COUNT' 변수**
mold_code 변경 시 count 변수 값들이
'1'로 리셋
제품 공정 시 마다 count 값 하나씩 추가

② 싸이클 분리 위해
trial 열 추가

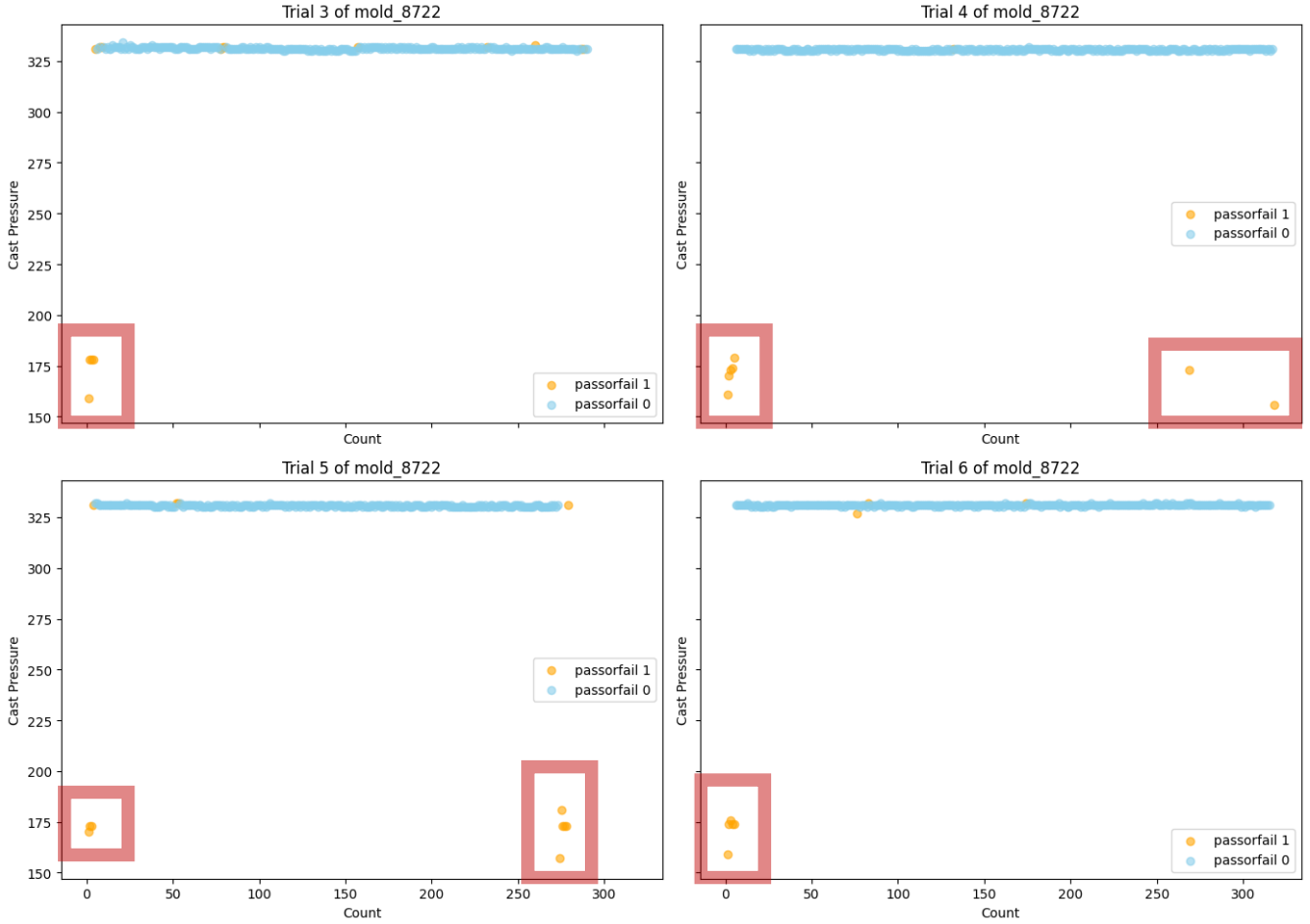
Count 값이 1로
리셋 될 경우
새로운 trial 값 생성

③ 'mold_code'변수 별 데이터 분리



'count' 변수의 초반 **불량품 집중** 확인

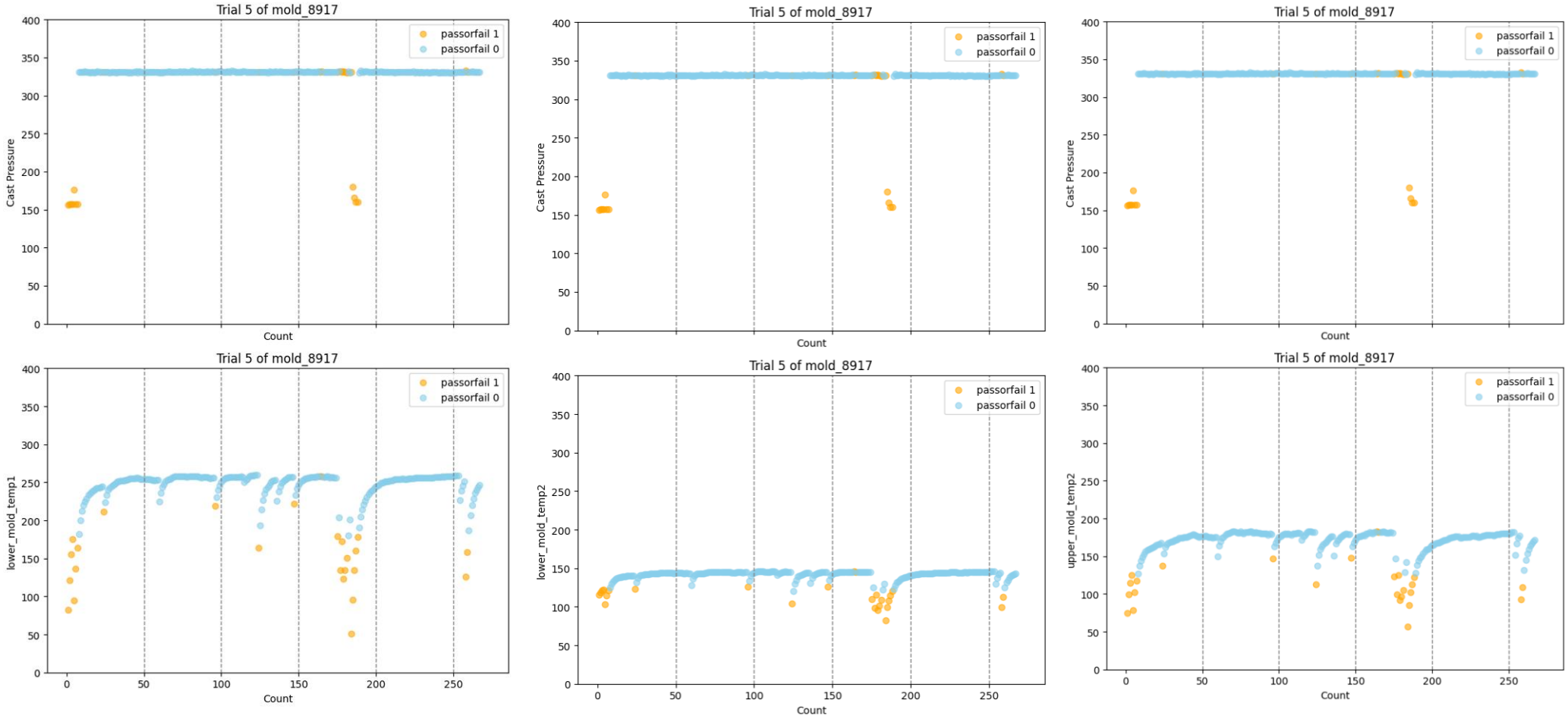
단위 공정 초반 불량



cast _ pressure 값이 작은 경우 불량 발생

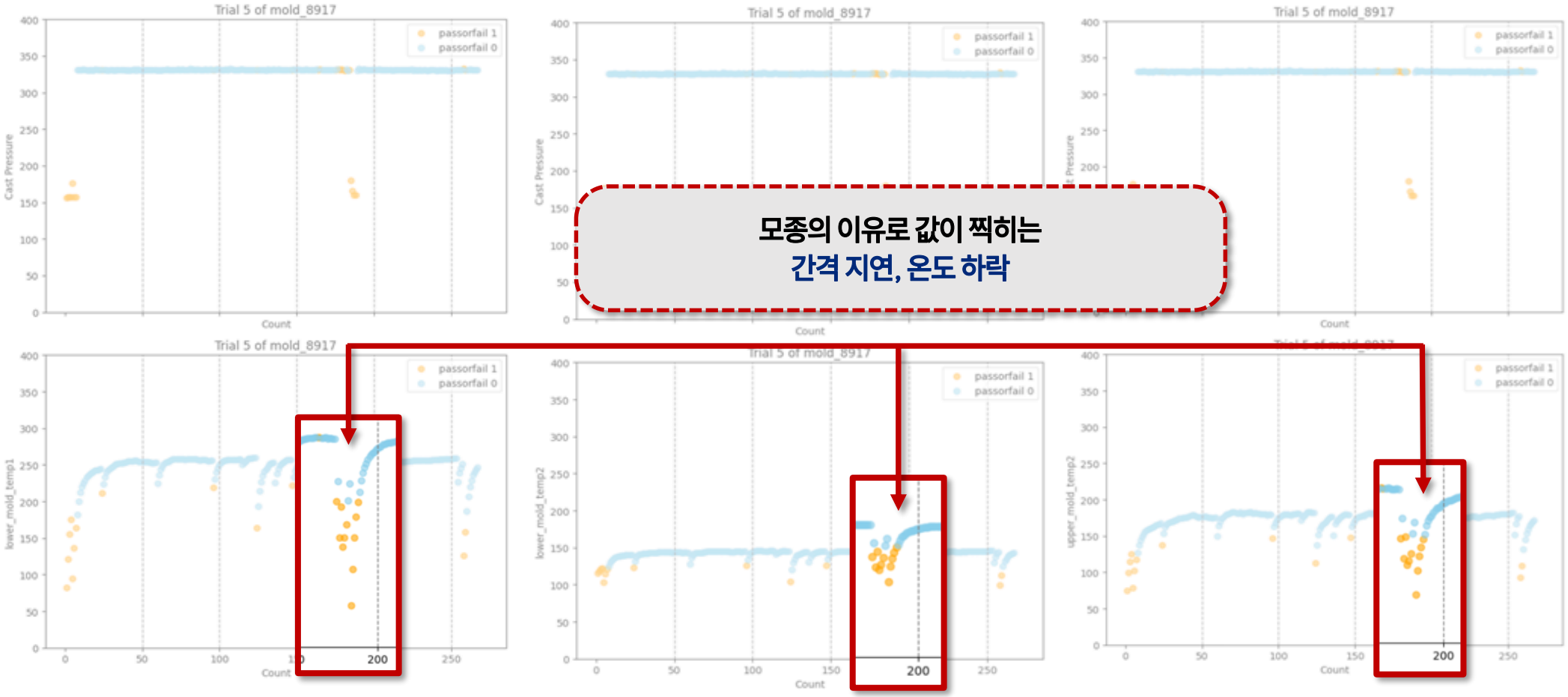
count 변수의 초반 불량 발생 경향

cast_pressure 하락 이유



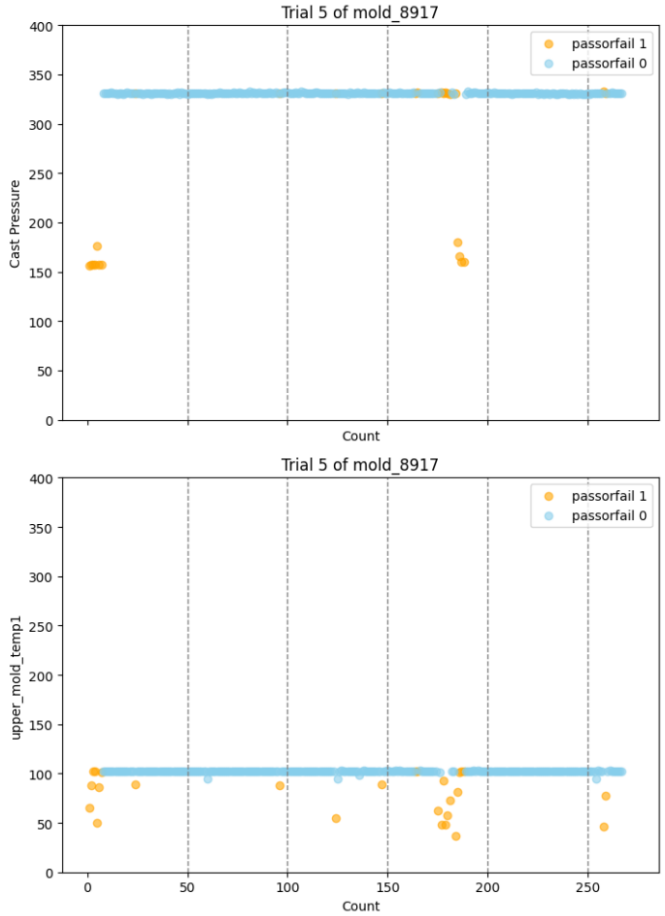
간헐적으로 lower_mold_temp1,2/upper_mold_temp2 의
온도가 **급격히 하락** 후 다시 **서서히 상승**하는 현상 발견

cast_pressure 하락 이유



2분 간격으로 count 값 연속적 생성 → 3분~4분 이상으로 벌어질 경우
lower_mold_temp1,2와 upper_mold_temp2 의 **급격한 하락**

cast_pressure 하락 이유



upper_mold_temp1 또한 온도가 급격히 떨어지지 않지만,
점진적인 하락이 관찰됨

cast_pressure 하락 이유

* $PV=nRT$

이상기체상태방정식

P는 기체의 압력, V는 기체의 부피
n은 기체의 몰 수, R은 기체 상수로 [약 8.314 J/(mol·K)]
T는 기체의 절대 온도(K)

▼

온도인 T 가 감소할 때, 압력인 P 도 감소	압력인 P 가 감소할 때, 온도인 T 도 감소

▼

공정에서 온도의 감소가 압력의 감소를 초래하는지 혹은 압력의 감소가 온도의 감소를 유발하는지

명확하지 않으나, **두 변수 중 하나가 감소하면 다른 변수도 감소하는 상호 의존적 관계가 있다고 판단**

시간 간격을 기록하는 *gap과 *gap_sign 파생 변수

제품 하나를 생산 시 평균적으로 2분 정도 소요되나, 가열로가 변경되거나 **싸이클(trial)이 변경되는 경우**에는 3~4분 이상의 시간이 소요
→ 시간 차이가 170초 이상이면 gap_sign=1 입력

경우 1) 가열로 변경으로 인한 시간 소요 (3~4분)

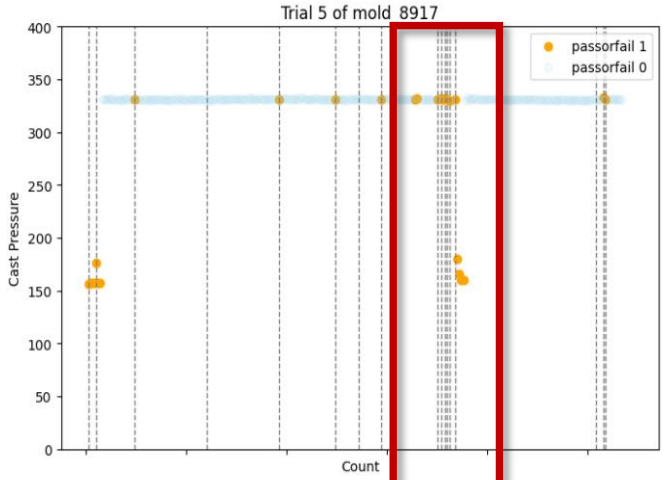
time	passorfail	heating_furnace	gap	gap_sign
13:14:48	0	B	128	0
13:16:46	0	B	118	0
13:20:29	1	A	223	1
13:22:34	0	A	125	0
13:24:32	0	A	118	0

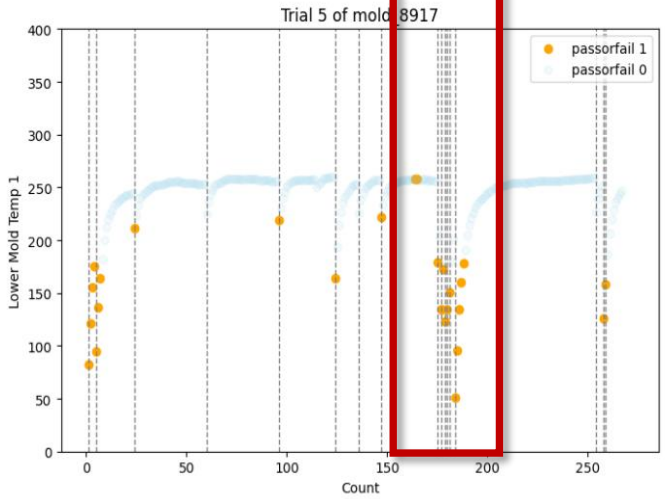
경우 2) trial의 변경(mold 변경/재사용 및 가동 중지로 인한 시간 차이 발생)

time	passorfail	heating_furnace	trial	count	gap	gap_sign
18:58:29	0	B	0	274	108	0
19:00:32	0	B	0	275	123	0
20:02:23	1	B	1	1	3711	1
20:05:22	1	B	1	2	179	1
20:07:24	1	B	1	3	122	0

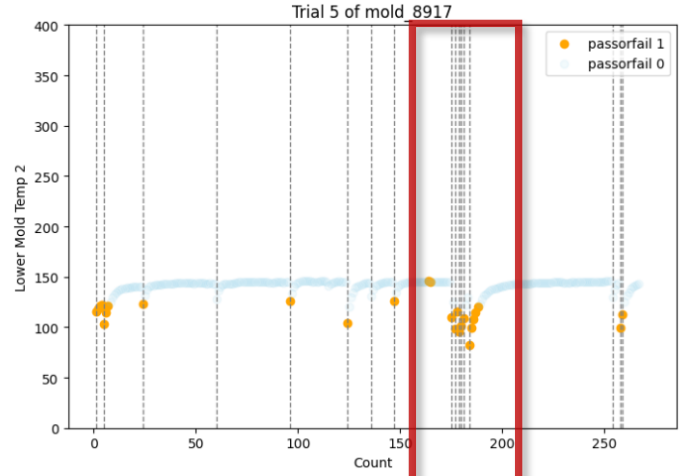
시간간격을 기록하는 *gap과 *gap_sign 파생 변수

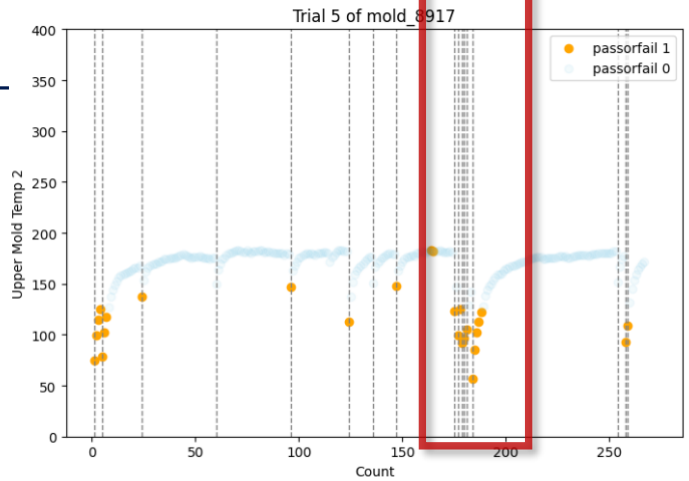
파생 변수	gap	약 2분 간격이 아닌 3~4분의 이상일 경우, 온도가 떨어지는 경향 발견 시간간격을 기록하는 'gap'열 추가	gap_sign	3~4분 이상일 경우, 시간차이 170초 이상 시 1, 아니면 0을 기록





gap_sign이 1일 때마다 세로선을 그었으며,
gap_sign이 1일 때마다 온도가 감소하는 현상이 관찰





결측치 확인

결측치 제거 근거

속성(column)	결측치 개수(개)
molten_temp	2261
molten_volume	45130
upper_mold_temp3	312
lower_mold_temp3	312
tryshot_signal	90095
heating_furnance	49145

거의 모든 열에서 공통적으로 결측치가 존재하는 행이 1개 존재 (19327번째 행) – 삭제

▼

결측치 개수 확인 (표 참고)

▼

결측치가 50% 혹은 그 이상인 열
['molten_volume', 'tryshot_signal', 'heating_furnace'] → **세 가지 열 제거**

working	
가동	91963
정지	51

'working' = '가동' 값이 전체의 99.945%
(하나의 동일한 값을 가지는 열 –제거)

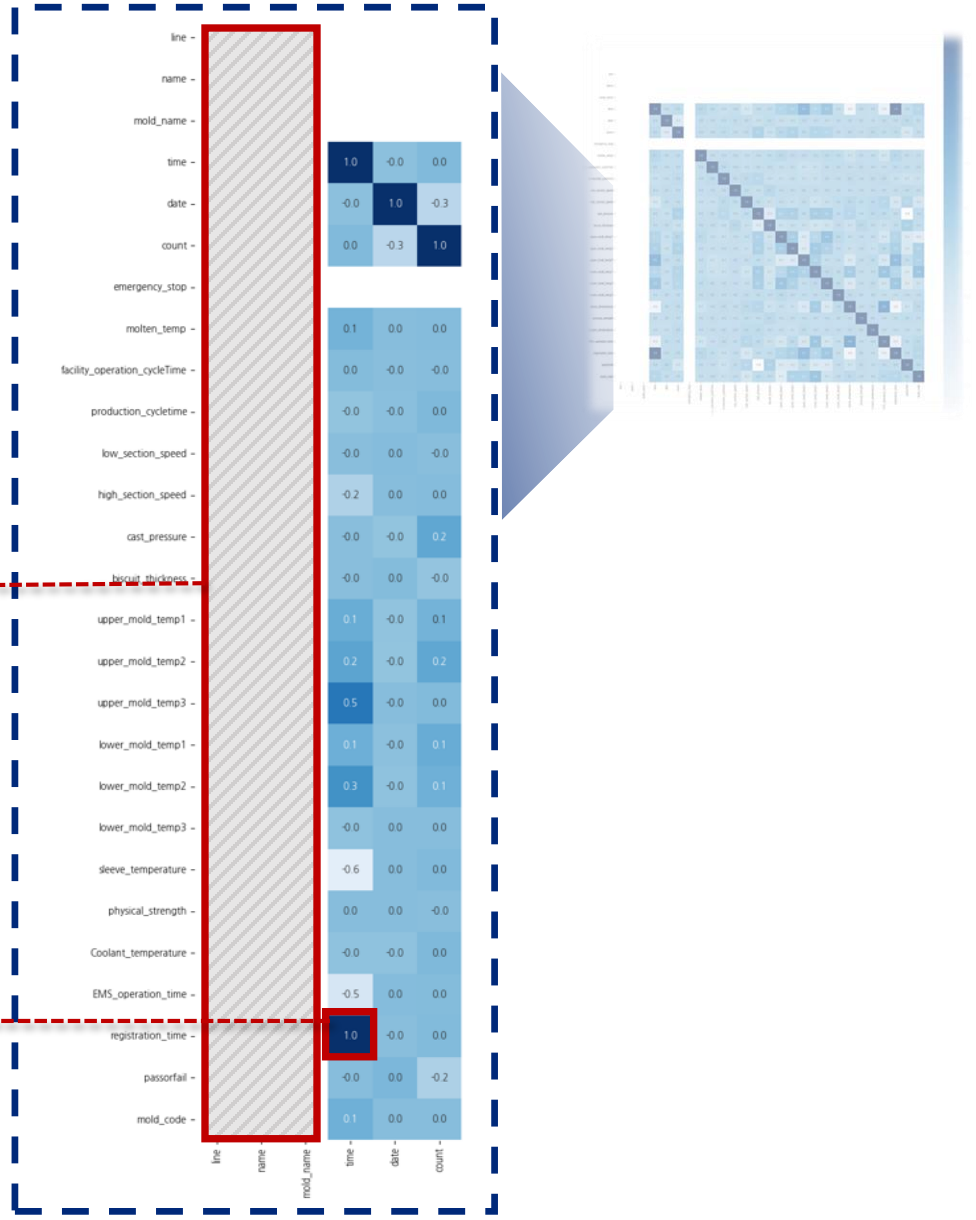
결측치 확인

결측치 제거 근거

데이터 타입이 object인 것들을 라벨 인코딩

변수 'line', 'name', 'mold_name', 'emergency_stop'
하나의 특정한 값만 가짐 – 해당 변수 제거

변수 'time' 과 'registration_time' 의 상관계수 : 1
두 변수가 동일함 – 해당 변수 제거



결측치 확인

결측치 제거 근거

F1-SCORE

평균 0.88610	중앙값 0.88814	Knn 0.88682	삭제 0.90216

molten_temp, upper_mold_temp3, lower_mold_temp3 변수의 결측치 값을 여러 방법으로 대체 혹은 삭제

▼

결측치를 삭제하였을 경우 F1-SCORE가 0.90216으로 가장 높음

▼ ▼

수치형 변수 제외 나머지 변수 삭제

범주형 변수 'count' 삭제

모델 학습 및 비교

데이터 불균형

양품 개수 87998

불량 개수 4016

불량률 4.3644553

개수 불균형 문제의 해결을 위한 StratifiedShuffleSplit

▼

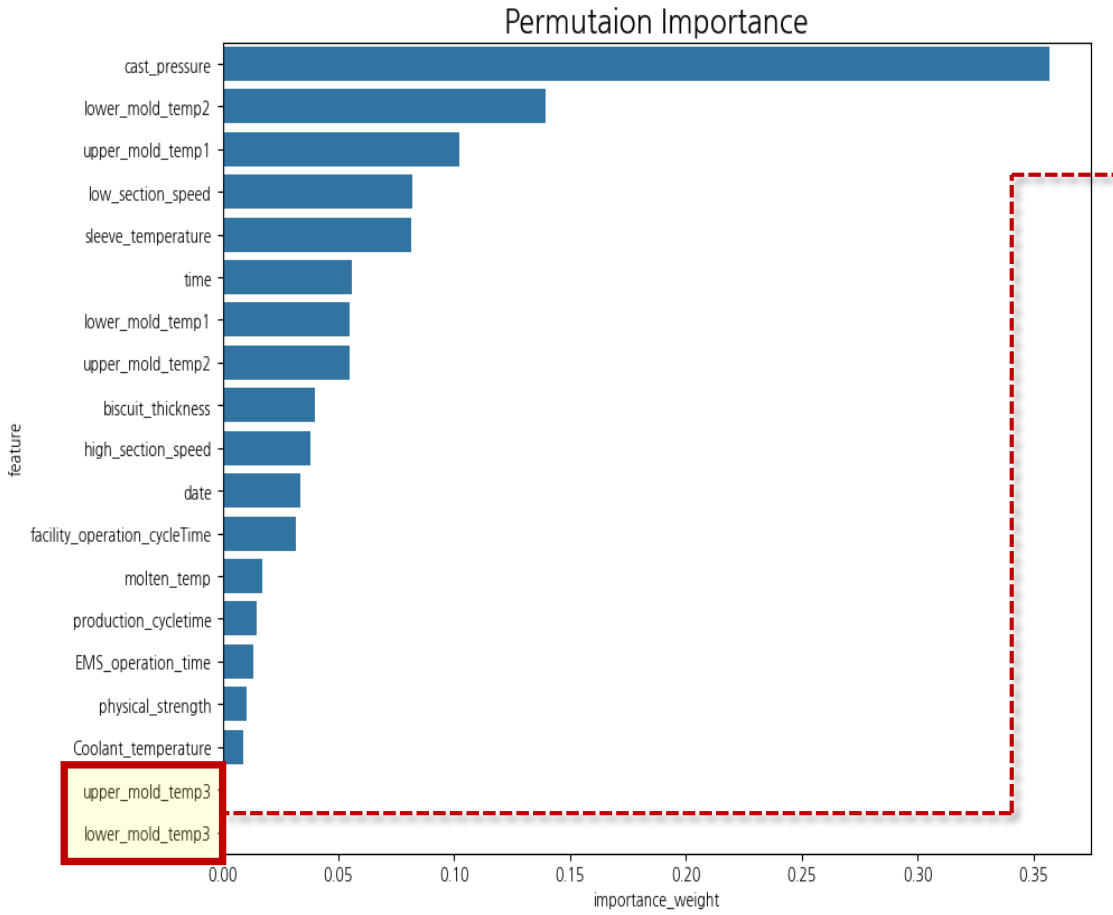
모델링을 위한 minmaxscaler 를 통한 정규화

▼

사용 모델 : 의사결정나무, 랜덤 포레스트, XGB, LGBM, ADA Boost

의사결정나무	랜덤포레스트	XGB	LGBM	ADA Boost
0.88987	0.91125	0.91870	0.91856	0.82989
		가장 높은 성능		

Permutation Importance



upper_mold_temp3, lower_mold_temp3 영향 X 변수

→ 해당 두 가지 변수 제거

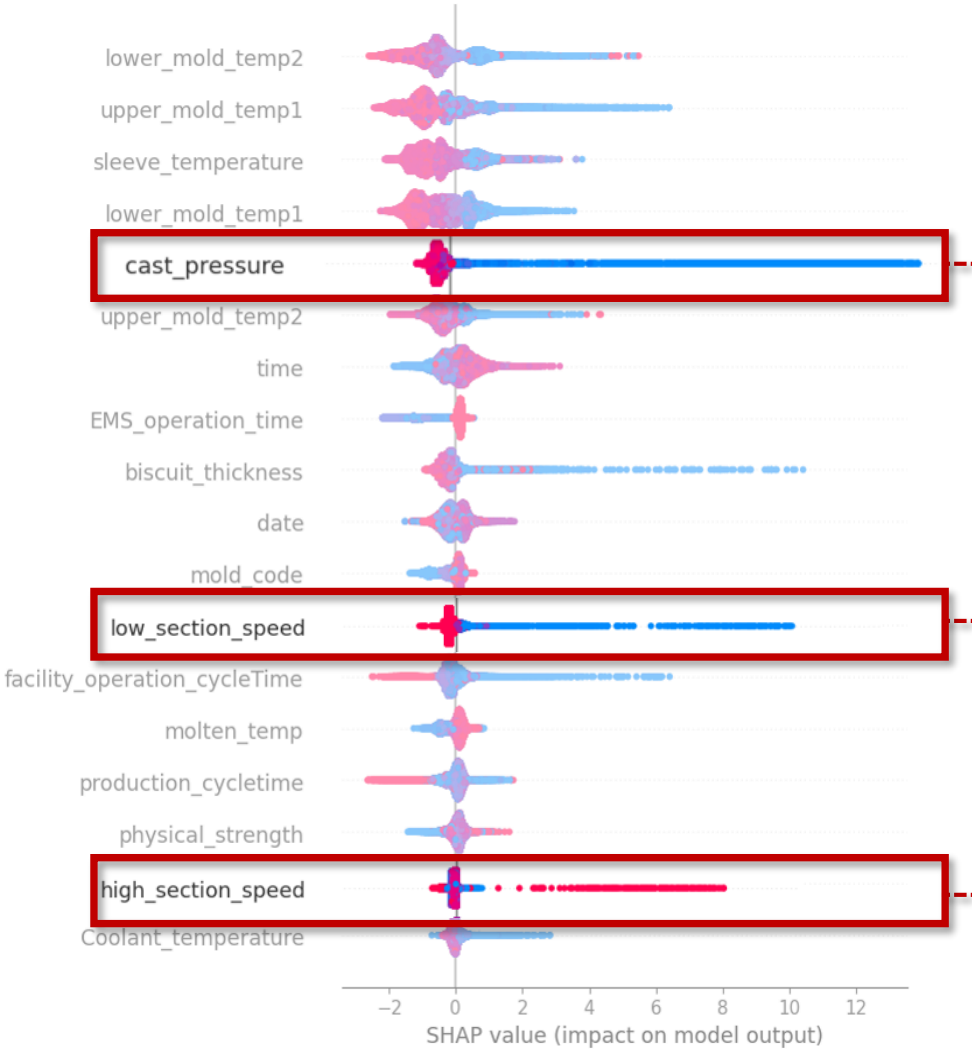
의사결정나무	랜덤포레스트	XGB	LGBM	ADA Boost
0.88483	0.91923	0.92238	0.91601	0.83109
		가장 높은 성능		

두 가지 변수 제거 전 (0.91870) 보다 성능 향상

모델 선택

	세부 내용		의사결정나무	랜덤포레스트	XGB	LGBM	ADA Boost
시도 1	이상치 제거	O	0.84899	0.89084	**0.90277**	0.89354	0.80721
	['EMS', 'mold_code'] 변수 제거						
시도 4	이상치 제거	X	0.87634	0.90861	0.91099	**0.91193**	0.82170
	['EMS', 'mold_code'] 변수 제거						
시도 5	이상치 제거	O	0.85038	0.89436	**0.89450**	0.88966	0.80798
	['EMS', 'mold_code'] 변수 제거						
시도 8	이상치 제거	X	0.87293	0.90656	0.90718	**0.90766**	0.82352
	['EMS', 'mold_code'] 변수 제거						
	가장 중요도가 낮은 변수 2개 제거						
시도 10	이상치 제거	O	0.84883	0.89124	**0.89921**	0.89413	0.80903
	['EMS', 'mold_code'] 변수 제거						
	가장 중요도가 낮은 변수 2개 제거						
시도 14	이상치 제거	X	0.88483	0.91923	**0.92238**	0.91601	0.83109
	['EMS', 'mold_code'] 변수 제거						
	가장 중요도가 낮은 변수 2개 제거						

XGBoost 모델에 따른 SHAP



변수 'cast_pressure'
'low_section_speed'
'high_section_speed'

중요도가 가장 높음

모델링 결론

① 이상치 정의 : 변수의 분포상 비정상적으로 극단적인 값을 가져
일반적으로 생각할 수 있는 범위를 벗어난 관측치

이상치는 이상치 일 뿐, 쓸모 없는 데이터가 아님
(실제로 이상치를 제거하지 않은 경우에 성능 향상)

② 다양한 데이터 전처리 시도 중요

③ 가이드라인에서는 mold _ code 변수를 제거하였으나,
EDA를 통해 해당 변수의 중요성을 발견하여 추가

이상치 처리 x + 결측치 제거 + 중요도가 낮은 변수 2가지 제거 + mode _ code 추가
Xgboost 모델
→ F1 SCORE 0.92238
향후에 데이터가 많아지면 mold _ code 별로 따로 모델링 제안

Dear. Factory Manager

공장장님께 드리는 편지…

① 공정 시작 초반에 불량이 발생하는 경향 발견. 이때, cast_pressure와 lower_mold_temp1,2와 upper_mold_temp2가 기준값 보다 낮은 경향 관찰.

솔루션 :
공정을 시작할때, cast_pressure와 lower_mold_temp1,2와 upper_mold_temp2가 기준값에 도달하지 않아 발생하는 가능성이 있어보이므로, 환경을 충분히 세팅한 후에 공정을 시작해야합니다.

② 가열로를 변경하면서 시간 간격이 벌어질 때, 불량이 발생하는 경향이 있다.

솔루션 :
1) 가열로를 변경할 때, 시간간격이 벌어지지않도록 공정 최적화
2) 가열로를 변경할 때 시간간격이 벌어지는 것이 필치 못하다면, cast_pressure와 lower_mold_temp1,2와 upper_mold_temp2가 기준값에 도달할 수 있도록 환경을 충분히 세팅한 후에 공정을 시작해야합니다.

③ 가열로를 변경하지 않더라도 시간 간격이 벌어지면, 불량이 발생하는 경향이 있다.

요구사항 : 필치 못한 경우가 아니라면 시간간격이 2분 정도를 유지할 수 있도록 해야합니다.