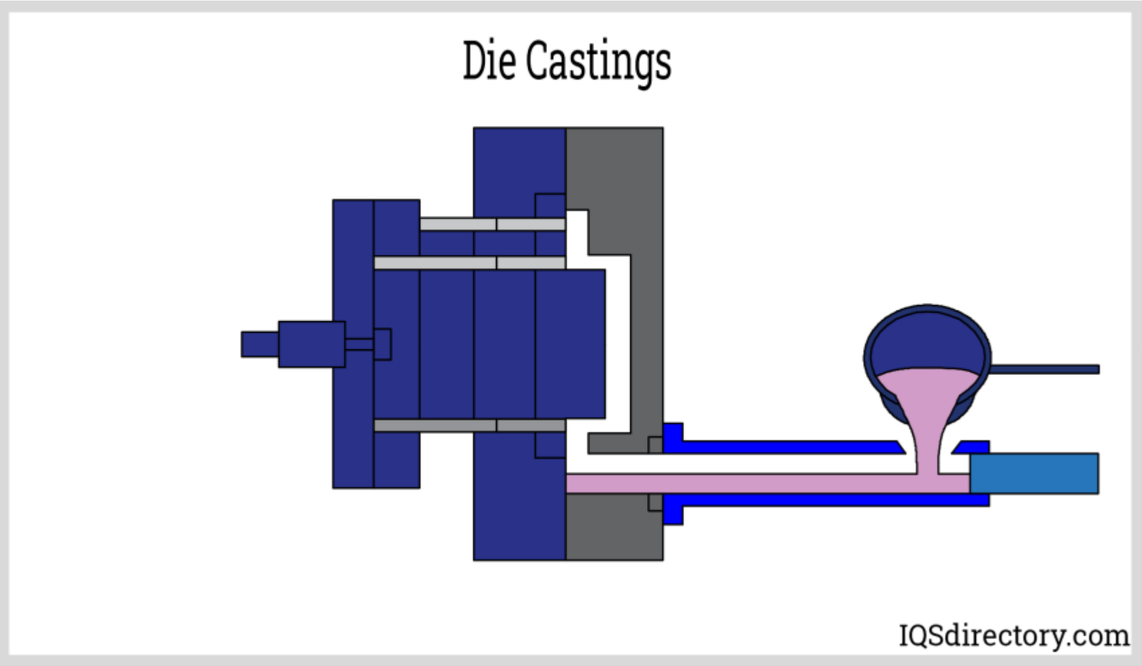
분석 주제 선정 보고서

4주차 다이캐스팅 실습 프로젝트

LS Big Data School Group 4
김서정, 박성민, 서성호, 윤주영, 이채원, 임유빈

CONTENTS

다이캐스팅(Die Castings)



- 온도(Temperature)
- 압력(Pressure)
- 속도(Velocity)
- 시간(Time)

다이캐스팅은 액체화된 금속을 주조(틀, Frame)에 넣고 원하는 모양의 금속부품을 생산하는 방법이다.

프로젝트 주제 및 선정배경

1. 문제상황

현장에서 온 정보

1	문제 1	데이터 제공 기업의 경우 일일 또는 주간 단위로 품질 이슈 현황을 파악하고 있으며 불량원인을 수작업으로 분석하고 있다.
2	문제 2	각 불량에 대한 발생원인과 대책이 정의되어 있으나 이를 적용하여 해결하지 못하고 있는 실정이다.
3	문제 3	대부분의 중소기업에서는 관리자 및 작업자의 경험에 의해 설비를 운용하고 있어 체계적인 관리를 하지 못한다.

일정한 공정 환경 및 공정 변수 관리 통해 불량에 대응하는 것이 필요!

데이터 탐색

1. 제조데이터 소개

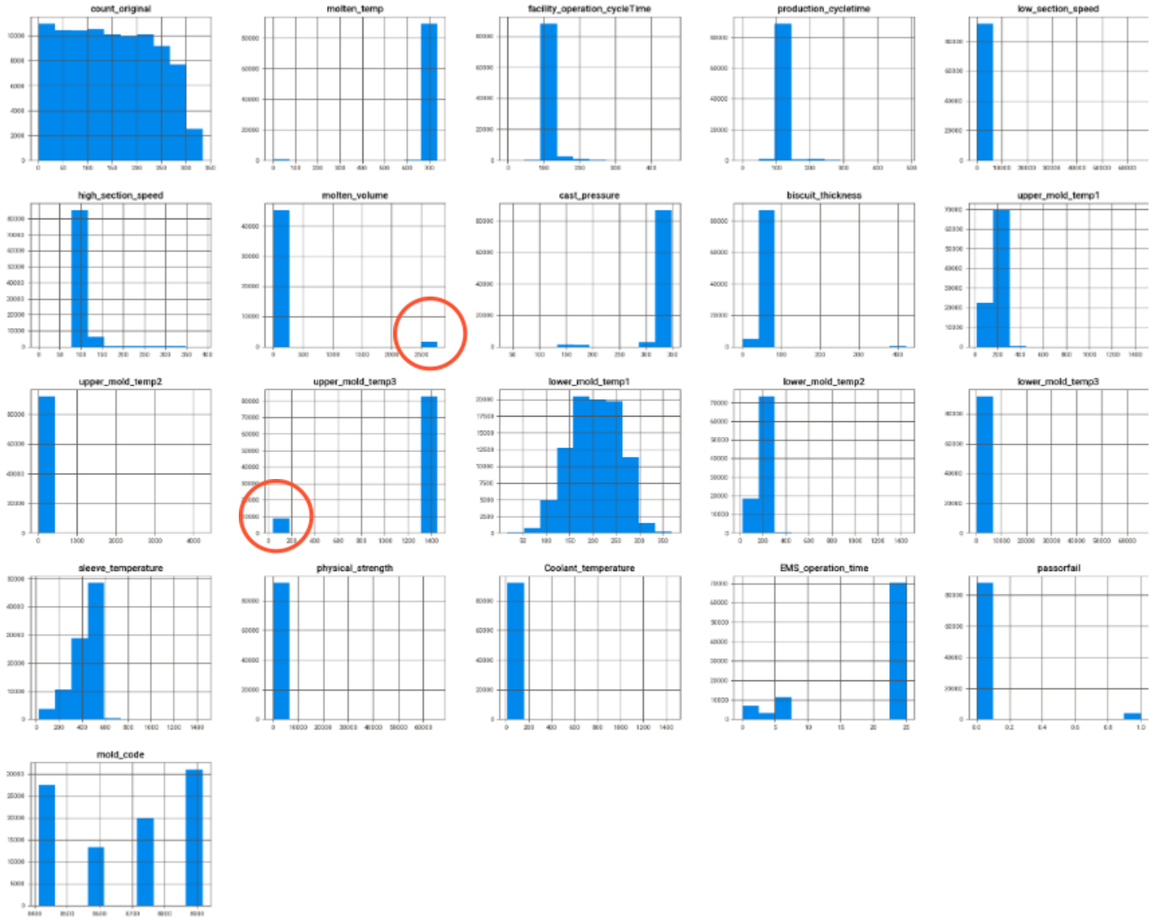
구분	명칭
독립변수	molten_temp
	production_CycleTime
	low_section_speed
	high_section_speed
	cast_pressure
	biscuit_thickness
	upper_mold_temp1
	upper_mold_temp2
	upper_mold_temp3
	lower_mold_temp1
	lower_mold_temp2
	lower_mold_temp3
	sleeve_temperature
	physical_strength
	Coolant_temperature
종속변수	passorfail

- 데이터 수집 방법
  - 주조 분야 : 다이캐스팅
  - 수집장비 : 주조 설비 내 PLC
  - 수집 기간 : 2019년 01월 02일 ~ 2019년 03월 31일
- 데이터 유형/구조
  - 데이터셋 구조 : 테이블 형식
  - 데이터 개수 : 총 2,852,465개(row 92,015개, column 31개)

- 변수 유형
  - object : line, name, mold_name, time, date, working, emergency_stop, registration_time
  - int64 : count, facility_operation_CycleTime, production_Cycletime, EMS_operation time, mold_code
  - float64 : molten_temp, low_section_speed, high_section_speed, molten_volume, cast_pressure, biscuit_thickness, upper_mold_temp1-3, lower_mold_temp1-3, sleeve_temperature, physical_strength, Coolant_temperature, passorfail

데이터 탐색

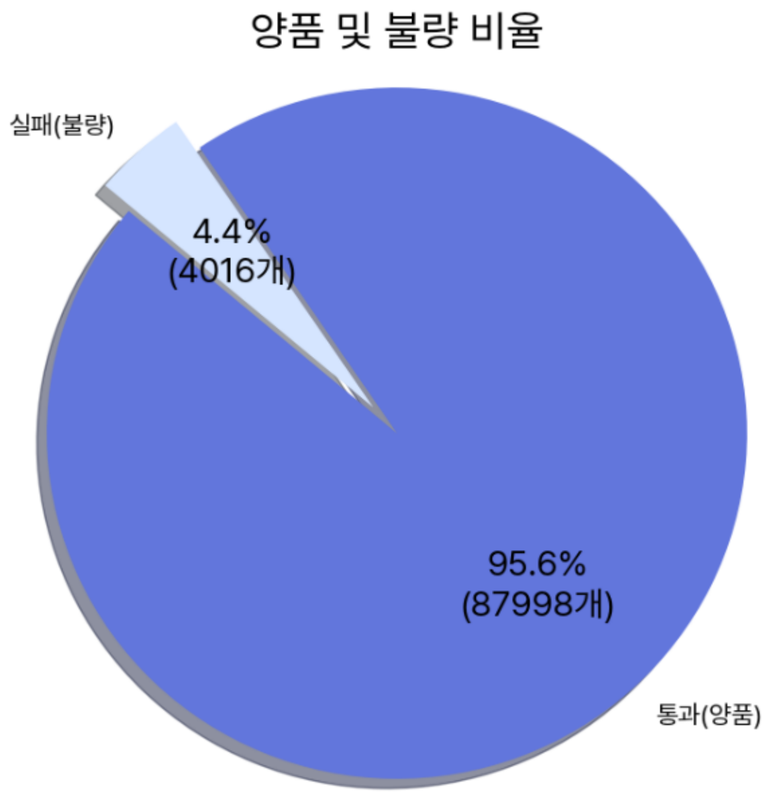
2. 히스토그램



- 히스토그램으로 데이터 값 분포 확인
- 본 시각화를 통해 정규분포에서 크게 벗어난 경우 이상치 존재여부를 확인할 수 있다!

데이터 전처리 및 시각화

1. 양품 및 불량 개수 확인



2. 숫자형 변수만 사용하기

숫자형이 아닌 변수 → 학습 불가능
필요한 데이터가 숫자형이므로 'object' 타입이 아닌 변수로 데이터를 재구성한다.

	count_original	molten_temp	facility_operation_cycleTime	production_cycletime	low_section_speed	high_section_speed	molten_volume	cast_pressure	biscuit_thickness	upper_mold_temp1	...	upper_mold_temp3	lower_mold_temp1	lower_mold_temp2	lower_mold_temp3	sleeve_temperature	physical_strength	Coolant_temperature	EMS_operation_time	passorfail	mold_code
0	258	731.0	119	120	110.0	112.0	75.0	331.0	35.0	186.0	...	1449.0	234.0	376.0	1449.0	550.0	786.0	34.0	23	0.0	8722
1	243	720.0	98	125	109.0	108.0	NaN	389.0	48.0	250.0	...	NaN	288.0	183.0	NaN	481.0	0.0	30.0	25	0.0	8412
2	244	721.0	98	122	109.0	108.0	NaN	389.0	48.0	250.0	...	NaN	288.0	183.0	NaN	481.0	0.0	30.0	25	0.0	8412
3	245	721.0	100	125	112.0	108.0	NaN	388.0	49.0	268.0	...	NaN	211.0	179.0	NaN	483.0	0.0	30.0	25	0.0	8412
4	246	721.0	99	123	109.0	118.0	NaN	389.0	49.0	276.0	...	NaN	276.0	187.0	NaN	486.0	0.0	30.0	25	0.0	8412
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
92810	71	731.0	122	122	100.0	101.0	NaN	331.0	49.0	83.0	...	1449.0	231.0	261.0	1449.0	279.0	726.0	30.0	0	0.0	8917
92811	72	731.0	120	121	100.0	101.0	NaN	331.0	54.0	83.0	...	1449.0	232.0	261.0	1449.0	279.0	732.0	30.0	0	0.0	8917
92812	73	732.0	122	135	100.0	101.0	NaN	331.0	52.0	79.0	...	1449.0	225.0	195.0	1449.0	272.0	731.0	30.0	0	0.0	8917
92813	74	732.0	133	122	100.0	101.0	NaN	331.0	53.0	82.0	...	1449.0	227.0	199.0	1449.0	280.0	723.0	30.0	0	0.0	8917
92814	75	732.0	122	123	100.0	101.0	NaN	331.0	56.0	84.0	...	1449.0	228.0	200.0	1449.0	289.0	705.0	30.0	0	0.0	8917

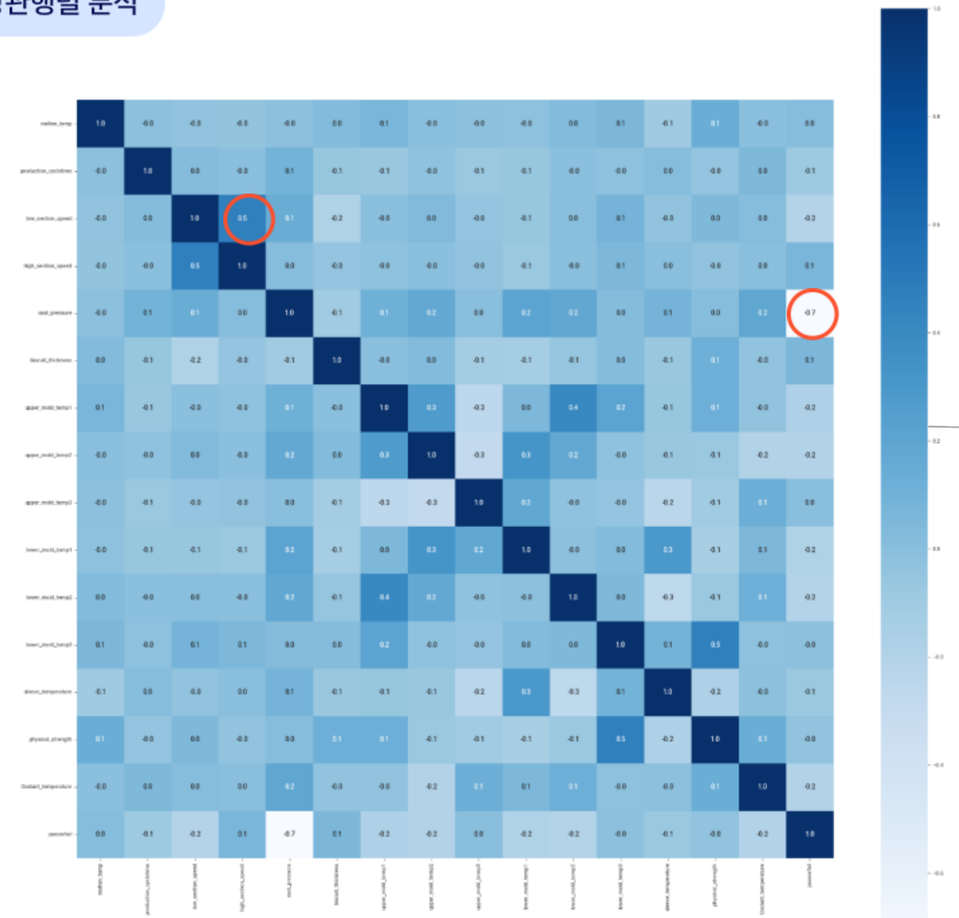
92015 rows × 21 columns

3. 데이터 탐색

데이터 사본 생성 → 칼럼 목록 확인 → 데이터프레임 크기 확인
→ 데이터프레임 null값 개수 확인 → 데이터프레임 통계 확인
→ 데이터프레임 정보 확인

데이터 전처리 및 시각화

4. 상관행렬 분석



Insight 1

passorfail과 cast_pressure의 상관계수는 0.7로 매우 높은 음의 상관관계를 보인다.

→ 주조압력이 낮은 값을 가질수록 불량품이 생산될 가능성이 높아진다!

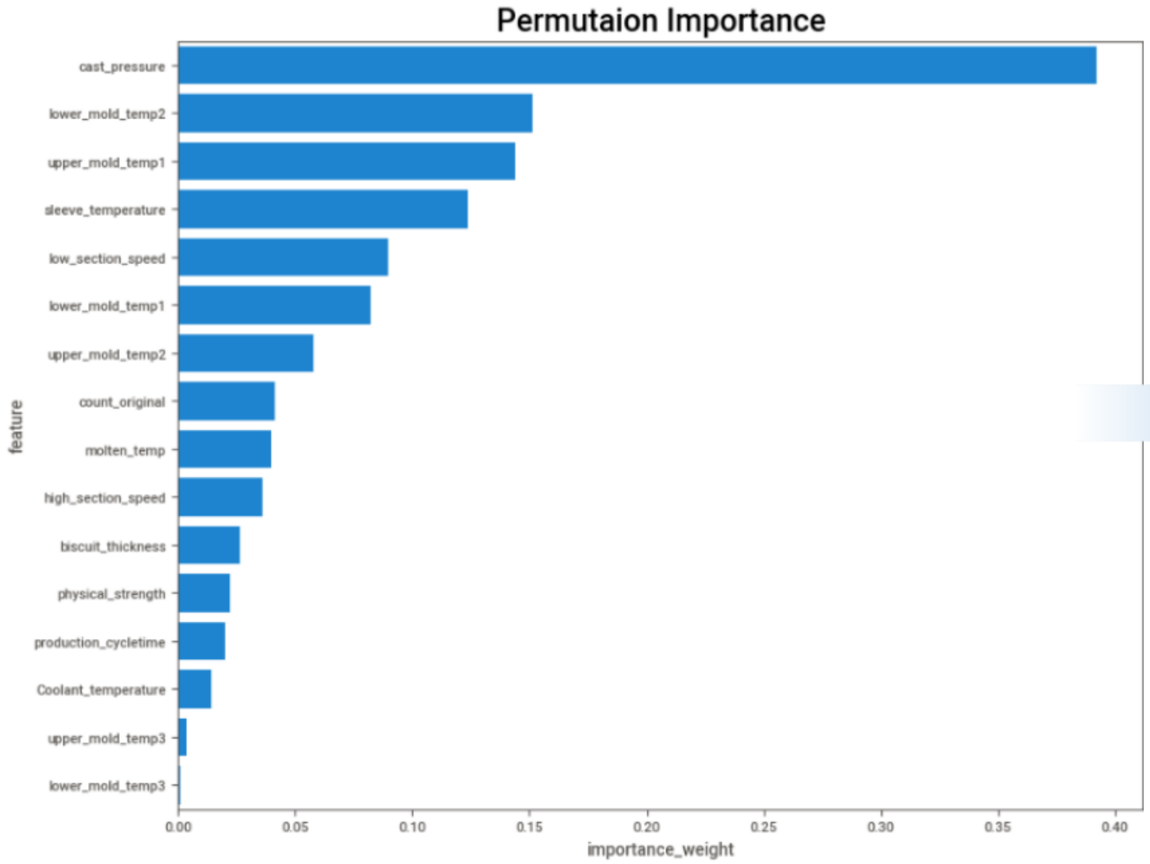
Insight 2

high_section_speed와 low_section_speed는 0.5로 높은 양의 상관관계를 보인다.

→ 저속구간속도와 고속구간속도가 주조 공정 내 밀접한 연관을 가진다!

결과 분석 및 해석

1. LightGBM 모델의 Permutation Importance 확인

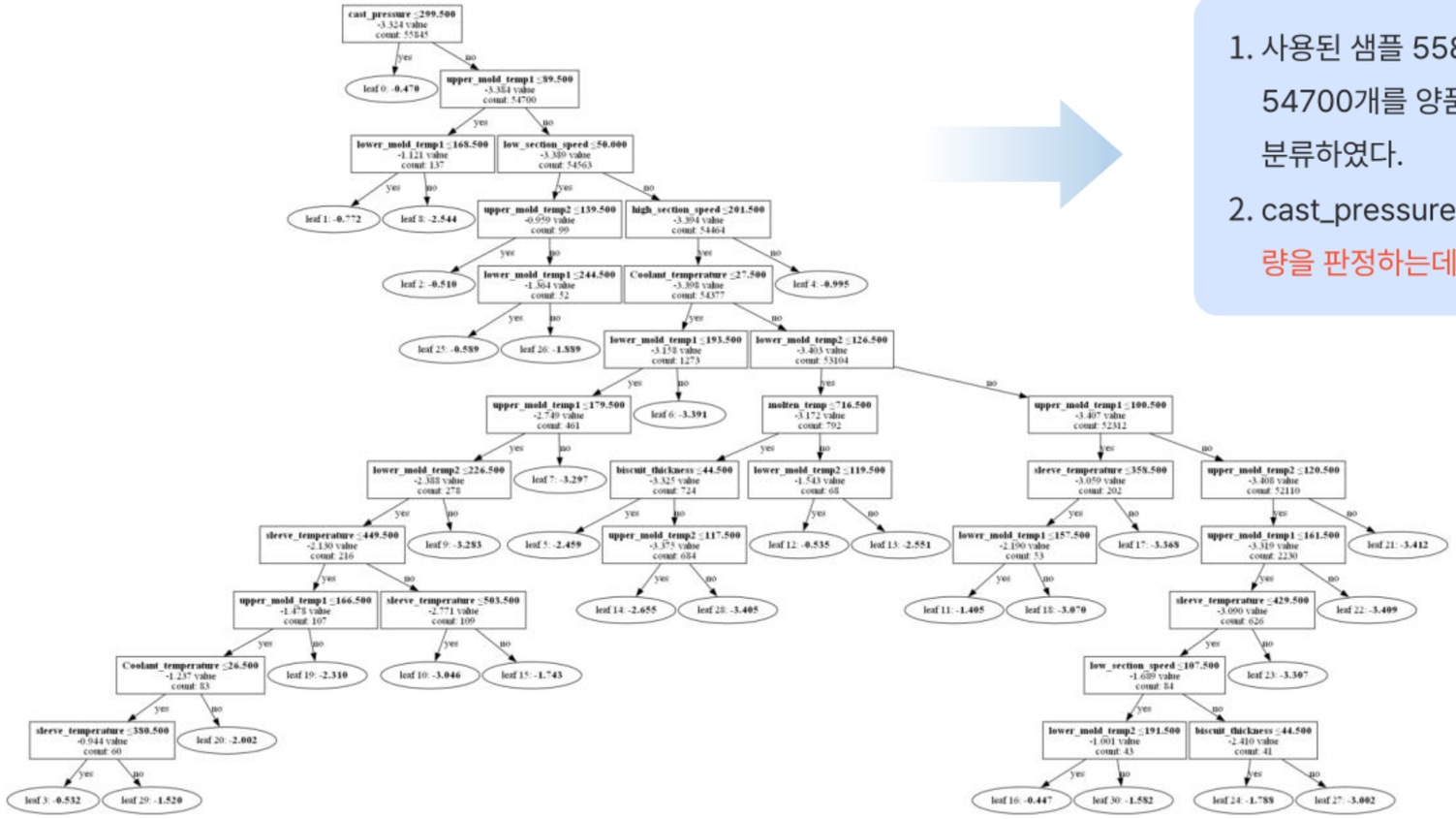


'cast_pressure'와 'lower_mold_temp2' 변수 중요도가 높다.

→ 해당 변수들이 주조 제품을 양품 or 불량으로 분류하는데 영향력이 높다.

결과 분석 및 해석

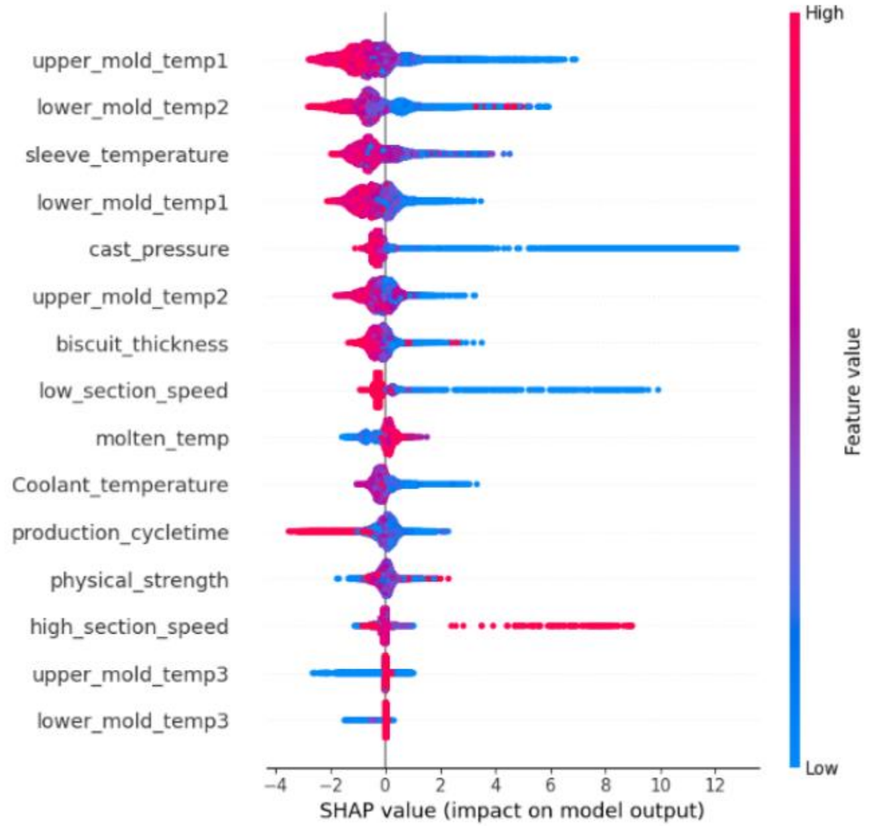
2. LightGBM 모델 트리 시각화 및 최적화



1. 사용된 샘플 55845개 중 cast_pressure = 299.5를 기준으로 54700개를 양품으로(>299.5), 1145개를 불량으로(<299.5) 분류하였다.
2. cast_pressure, upper_mold_tmp1 등 주요변수들이 양품과 불량을 판정하는데 영향을 미쳤음을 확인할 수 있다.

결과 분석 및 해석

3. SHAP 시각화



1. Shapley 값을 통해 각 특성이 예측에 기여하는 정도를 알 수 있다.
2. cast_pressure의 수치가 작은 값을 가질수록 불량률은 높아진다.
3. high_section_speed의 수치가 높은 값을 가질수록,
low_section_speed의 수치가 낮은 값을 가질수록 불량률은 높아진다.

인사이트 도출하기

1. 결측치 처리 및 이상치

결측치 및 이상치 처리 여부

INSIGHT 1

<기존 보고서>

50% 결측인 molten_volume칼럼은 제거하고 나머지 칼럼의 결측행을 제거하는 방식을 택함.
변수들에 존재하는 이상치를 상·하한 0.1% 해당하는 값으로 제거하였음.

<수정 방향>

☑결측치를 제거하는 대신 **평균값이나 보간법** 등을 활용하여 채운 뒤, 모델의 성능을 비교하여 최적의 결측치 처리 방법을 알아보고자 한다.
☑**이상치의 상·하한 설정 범위를 조정**해서 최적의 성능을 나타내는 범위를 결정하는 방법에 대해 알아보고자 한다.

2. F1 score 향상

F1 score 향상할 수 있는 방법

INSIGHT 2

<기존 보고서>

모델링을 통해 생성된 모델의 평균 F1 스코어가 가장 높았던 것은 'LightGBM'이다. 해당 모델의 F1 스코어는 split 1에서 0.895로 테스트 데이터에서 가장 높은 정확도를 보였다

<수정 방향>

☑AutoML을 통해 F1 스코어가 가장 높게 나온 모델은 **'XGBoost'**이다.
하이퍼파라미터 조정 등을 통해 **F1 스코어**를 향상시킬 수 있는 방법에 대해 알아보고자 한다.

보완할 점

자체 평가

감사합니다