2025.10.02

PRE-PROCESSING REPORT

데이터 전처리 상세 보고서

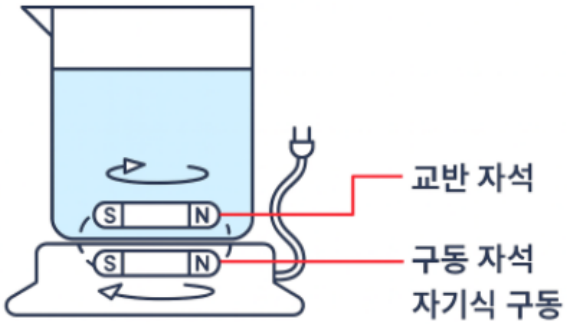
3조 좋은데요

CONTENTS

목차

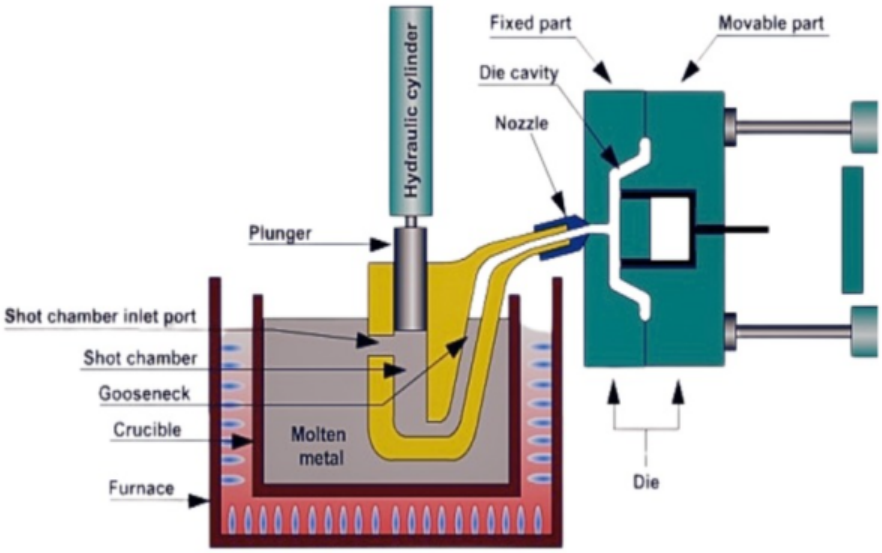
개요

공정 및 구성요소 설명



01 전자교반

- 금속 용융액을 부드럽게 회전시켜 열과 조성을 균일화



02 용광로 및 몰드

- 용광로에서 금속을 완전히 용해하고, 주입 전까지 안정적인 온도 유지
- 몰드로 금속의 모양을 잡아주고, 냉각시켜 금속을 응고시킴

VARIABLE

변수 설명

데이터의 변수 설명

01 전자교반(EMS) 변수

- 전자교반 가동 시간(EMS_operation_time): 전자교반을 가동하여 금속을 섞는 시간, [0, 3, 6, 23, 25]

02 용광로(Heating Furnace) 관련 변수

- 용탕 온도(molten_temp): 주입 전 용광로에서의 용탕 온도
- 용탕량(molten_volume): 용광로에 남아 있는 용탕량
- 용광로(heating_furnace): 용광로를 구분하는 변수, [A, B, nan]

03 몰드(Mold) 및 용탕 주입 관련 변수

- 상금형 온도(upper_mold_temp1~3): 상부 몰드의 온도
- 하금형 온도(lower_mold_temp1~3): 하부 몰드의 온도
- 냉각수 온도(Coolant_temperature): 몰드를 식히는 냉각수 온도
- 슬리브 온도(sleeve_temperature): 몰드에 용탕을 주입하는 주입부의 온도
- 저속 구간 속도(low_section_speed): 몰드에 용탕을 주입하기 시작할 때의 속도(저속)
- 고속 구간 속도(high_section_speed): 몰드에 용탕이 어느 정도 찼을 때 용탕을 주입하는 속도(고속)
- 주조 압력(cast_pressure): 피스톤이 몰드 안에 용탕을 주입할 때 걸리는 압력
- 비스켓 두께(biscuit_thickness): 몰드에 용탕 주입을 마친 뒤 주입 통로 쪽에 남는 금속 덩어리의 두께
- 형체력(physical_strength): 몰드에 용탕 주입 후 굳을 때까지 몰드를 고정하는 힘

변수 설명

데이터의 변수 설명

04 구분 및 메타 변수

- id: 각 행별 고유값, 식별용 변수
- 라인명(line): 작업 라인 설명, '전자교반 3라인 2호기'로 고정
- 제품명(name): 제품명, 'TM Carrier RH'로 고정
- 몰드 이름(mold_name): 몰드명, 'TM Carrier RH-Semi-Solid DIE-06'로 고정
- 몰드 코드(mold_code): 몰드 코드, [8722, 8412, 8573, 8917, 8600]
- 수집 일자(time): 데이터를 수집한 날짜(제품 생산 날짜), 2019-01-02부터 2019-03-12까지
- 수집 시간(date): 데이터를 수집한 시각(제품 생산 시각), 시:분:초 형식
- 등록 일시(registration_time): 데이터를 등록한 시각, 연-월-일 시:분:초 형식
- 생산 번호(count): mold_code별 수집 일자별 생산 번호
- 가동 여부(working): 설비 가동 여부, ['가동', '정지', nan]
- 트라이샷 신호(tryshot_signal): 시험 생산 여부, [nan, 'D'], 'D'일 경우 시험 생산
- 비상 정지(emergency_stop): 비상 정지 여부, ['ON', nan], 'ON'일 경우 정상, nan일 때 비상 정지
- 설비 작동 사이클 시간(facility_operation_cycleTime): 제품 생산 시 설비 작동 시간
- 제품 생산 사이클 시간(production_cycletime): 제품 생산 시 총 소요 시간
- 양품/불량 라벨(passorfail): 양품인지 불량인지 판별하는 라벨, 0.0일 경우 양품, 1.0일 경우 불량품

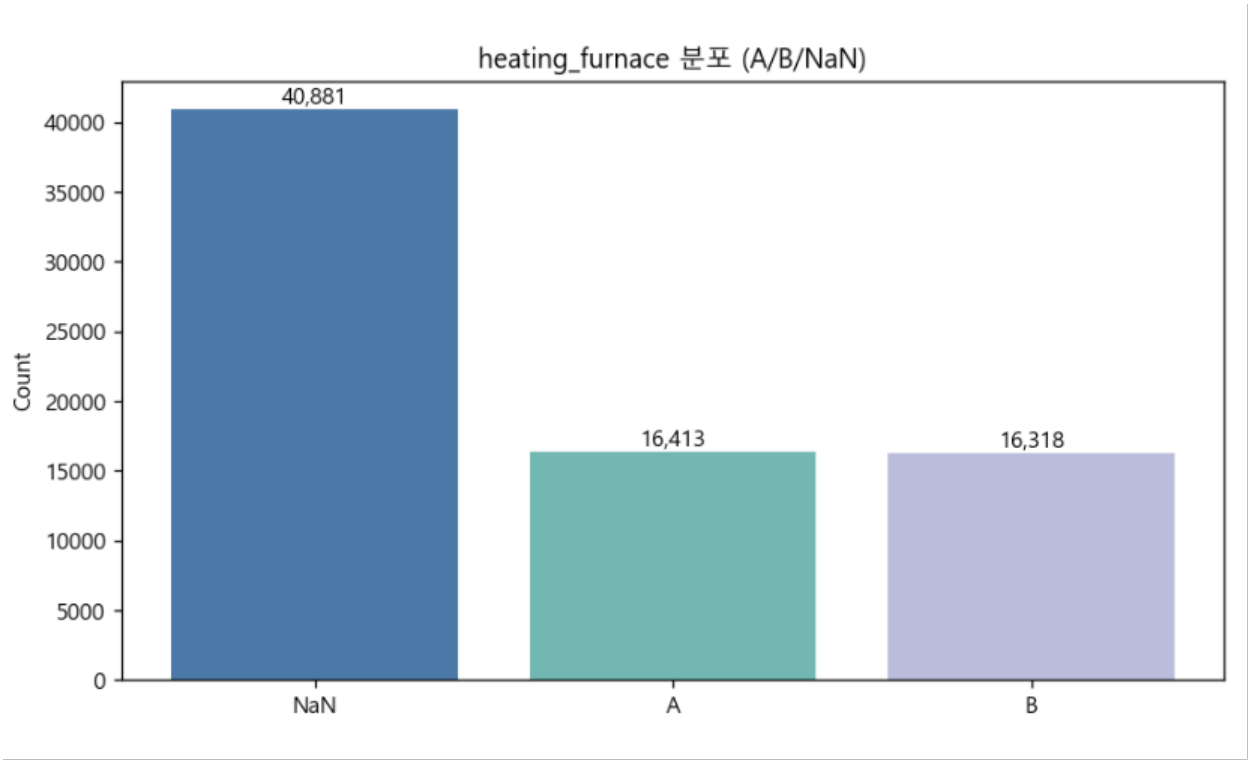
DROP COLUMN

단일 칼럼 제거

칼럼 전체를 제거한 경우

01 heating_furnace 열

- 전체 데이터 73,612개 중 결측치가 40,881개로 약 55.5%를 차지
- 'A'가 16,413개, 'B'가 16,318개인데 'NaN'이 40,881개로 'A', 'B'의 2배 이상
  ⇒ 'A', 'B'와 유사한 그룹 2개 이상으로 이루어졌다고 추측



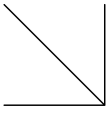
DROP COLUMN

단일 칼럼 제거

칼럼 전체를 제거한 경우

index	heating_furnace	mold_code	time	date	molten_volume	count
73406	B	8412	2019-03-12	03:45:28	Nan	204
73407	Nan	8917	2019-03-12	03:45:40	61.0	222
73408	Nan	8722	2019-03-12	03:46:35	84.0	219
73409	B	8412	2019-03-12	03:47:37	Nan	205
73410	Nan	8917	2019-03-12	03:47:46	60.0	223
73411	Nan	8722	2019-03-12	03:48:41	84.0	220

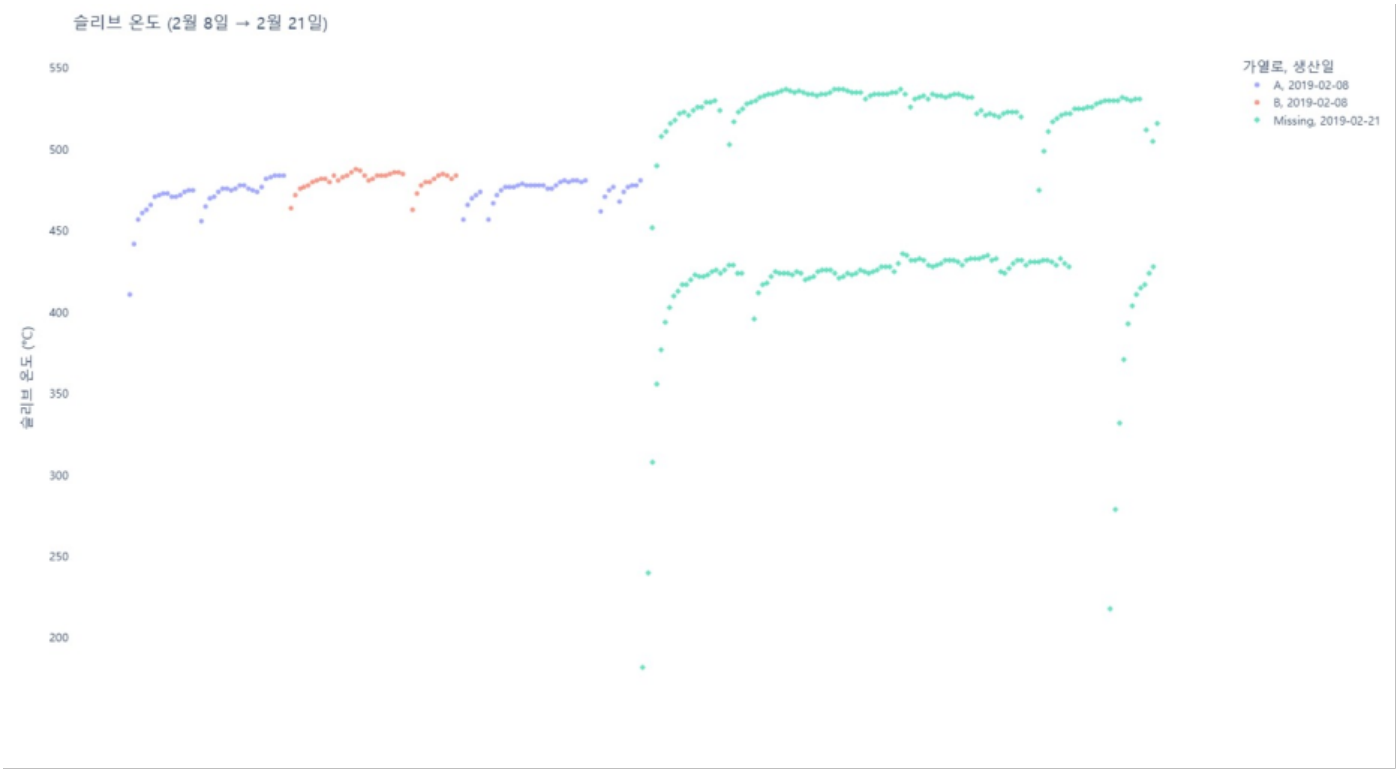
- 'A', 'B', 'nan' 집단을 비교해보면 규칙 확인 가능
- 결측이 아닌 구간: mold_code/time 일정, date/count 연속 ⇒ 동일 furnace에서의 연속 생산
- 연속 결측 구간(예: index 73407, 73408): mold_code 8917/8722로 상이, molten_volume 61.0→84.0로 불연속, count 222/219로 불연속 ⇒ 서로 다른 furnace로 보임
- 결측 구간 비교 1 (예: index 73407, 73410): mold_code 8917로 동일, molten_volume 61.0→60.0으로 연속, count 222→223으로 연속 ⇒ 동일 furnace에서의 연속 생산으로 해석 가능
- 결측 구간 비교 2 (예: index 73408, 73411): mold_code 8722로 동일, mlolten_volume 84.0으로 동일, count 219→220으로 연속 ⇒ 동일 furnace에서의 연속 생산으로 해석 가능
- 따라서 결측 구간은 최소 2개 이상의 상이한 집단(furnace)일 것이라고 예상 가능
- 두 집단이 A, B와 같은지 다른지 알 수 없고, mold_code를 바꾸거나 molten_volume을 새로 채울 때 furnace를 교체하는지 유지하는지 등의 규칙을 알 수도 없으므로 결측치를 채우기 어려움
- 기본 모델에서 heating_furnace의 변수 중요도도 높지 않음 ⇒ 최종적으로 heating_furnace 열 제거



DROP COLUMN

단일 칼럼 제거

칼럼 전체를 제거한 경우



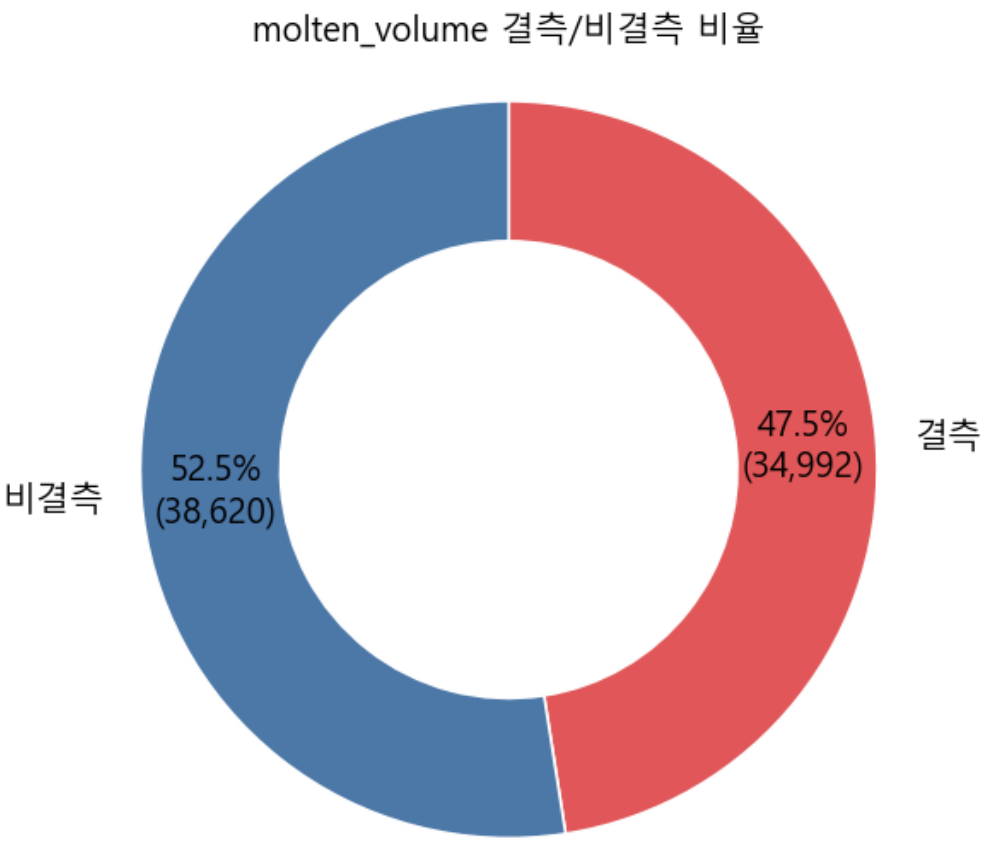
- 위 그래프(registration_time - sleeve_temperature 산점도)로 Nan이 상이한 집단임을 더 확실하게 볼 수 있음
- 파란색 점은 heating_furnace가 'A'인 경우, 빨간색 점은 heating_furnace가 'B'인 경우
- 초록색 점들이 heating_furnace 값이 결측인 경우 → 그러나 같은 시각에 두 개 패턴이 동시에 나타남
  ⇒ 결측인 경우도 최소 2개 이상의 집단으로 분류됨을 알 수 있음

단일 칼럼 제거

칼럼 전체를 제거한 경우

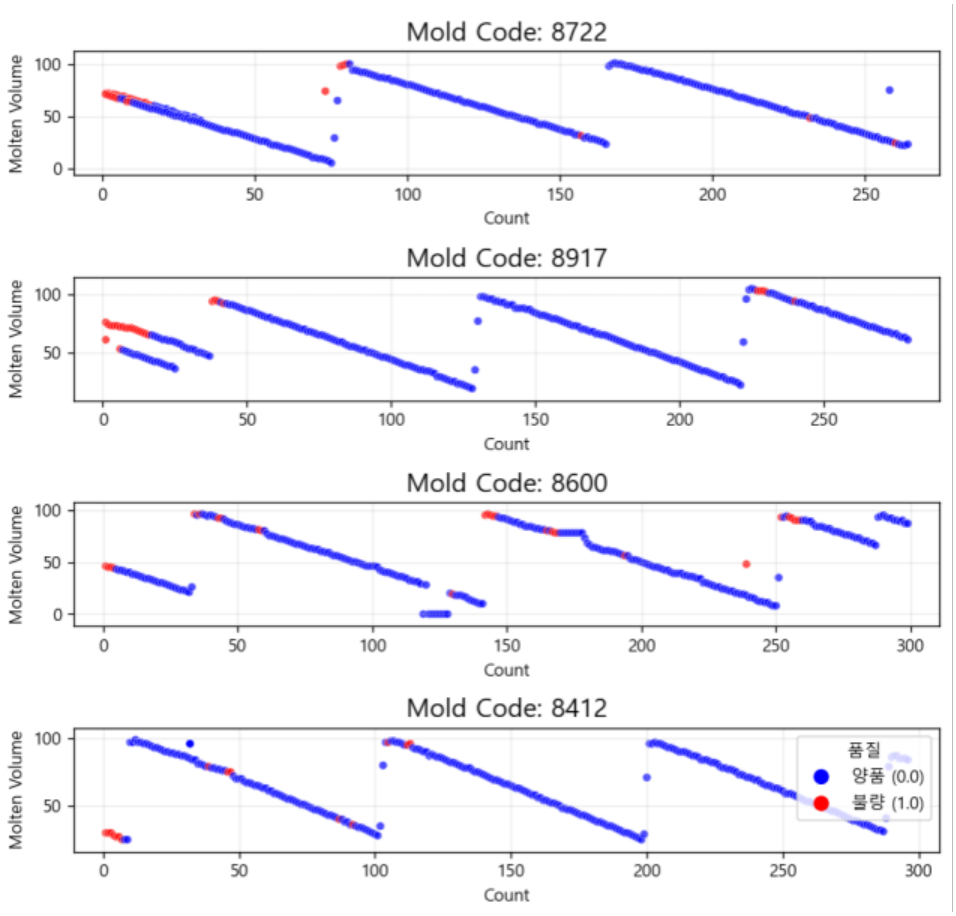
02 molten_volume 열

- 전체 데이터 중 결측치가 34,992개로 약 47.5%를 차지



단일 칼럼 제거

칼럼 전체를 제거한 경우



- mold_code별로 나눠서 count에 따라 molten_volume 그래프를 그렸을 때,
  count에 따라 molten_volume이 채워지고 다시 줄어드는 양상이 보임
- 따라서 결측치가 중간에 있는 경우 선형 회귀나 KNN 모델로 대치할 수 있다고 판단
  → 그러나 실제로는 결측이 연속되는 구간이 너무 길음
- molten_volume을 채우는 시간도 정해져 있지 않고 조금씩 달라져서 결측치를 채우기 어려움
- 기본 모델에서 molten_volume의 변수 중요도도 높지 않음 ⇒ 최종적으로 molten_volume 열 제거

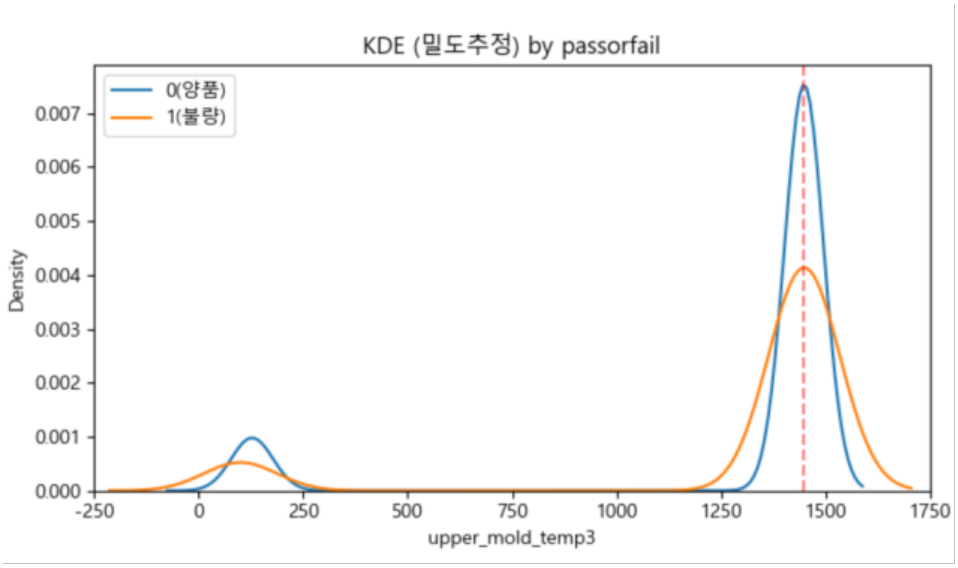
DROP COLUMN

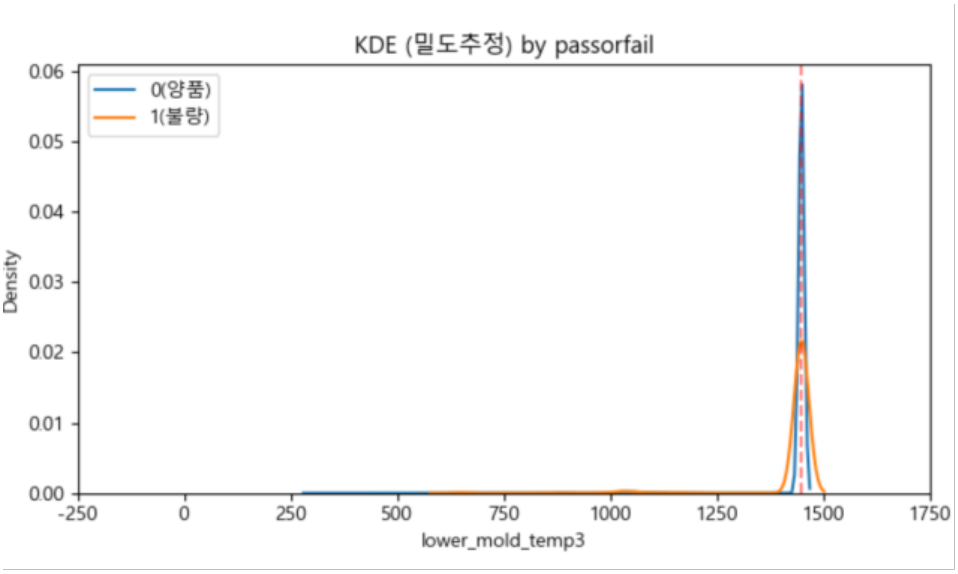
단일 칼럼 제거

칼럼 전체를 제거한 경우

03 upper/lower_mold_temp3 열

- upper_mold_temp3 열: 전체 데이터 73,299개 중 이상치(1449.0) 64,356개로 약 87.8% 차지
- lower_mold_temp3 열: 전체 데이터 73,298개 중 이상치(1449.0) 71,650개로 약 97.8% 차지
  → 이상치 1449.0을 센서 오류 코드로 가정 ⇒ 최종적으로 upper/lower_mold_temp3 열 제거





EXPLAIN DROPPING COLUMNS

단일 칼럼 제거

칼럼 전체를 제거한 경우

04 registration_time 열

- time 열: (데이터 등록 일자) 연-월-일 형식
- date 열: (데이터 등록 시각) 시:분:초 형식
- registration_time 열: (데이터 등록 일시) 연-월-일 시:분:초 형식
  → time 열과 date 열의 결합 정보로 의미 중복 ⇒ 최종적으로 registration_time 열 제거

registration_time	time	date
2019-01-02 16:45:06	2019-01-02	16:45:06
2019-01-02 16:45:08	2019-01-02	16:45:08
2019-01-02 16:45:58	2019-01-02	16:45:58
2019-01-02 16:48:03	2019-01-02	16:48:03
2019-01-02 16:50:08	2019-01-02	16:50:08

DROP ROW

행 제거

행 전체를 제거한 경우

01 emergency_stop

- emergency_stop이 결측인 경우 딱 1번 존재(train_df) → 이 행의 나머지 변수들 대부분 결측치
  ⇒ 모델 학습 데이터에서 emergency_stop이 결측인 행 제거
- 모델 예측 후 emergency_stop 값을 확인해서 결측인 경우 불량으로 판정하도록 함

id	time	date	count
19327	2019-01-25	19:09:09	281
working	emerygency_stop	facility_operation_cycleTime	production_cycletime
Nan	Nan	0	0
low_section_speed	high_section_speed	cast_pressure	biscuit_thickness
Nan	Nan	Nan	Nan
upper_mold_temp1	upper_mold_temp2	upper_mold_temp3	lower_mold_temp1
Nan	Nan	Nan	Nan
lower_mold_temp2	lower_mold_temp3	sleeve_temperature	physical_strength
Nan	Nan	Nan	Nan
Coolant_temperature	EMS_operation_time	tryshot_signal	heating_furnace
Nan	0	<NA>	Nan

DROP ROW

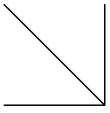
행 제거

행 전체를 제거한 경우

02 count 중복

- count, mold_code, time, molten_volume 등이 일치하는 경우 다른 모든 변수들도 같은 값을 가짐
- 특히 molten_volume이 변하지 않은 경우 용탕을 전혀 소모하지 않았다는 뜻 → 생산이 이루어지지 않음
  ⇒ 정보 중복을 피하기 위해 제일 앞의 한 행만 남기고 나머지 중복 행들은 삭제

id	time	date	count	mold_code	molten_volume	...	working	emergency_stop	passor fail
2953	2019-01-07	19:21:55	32	8412	2767.0	...	가동	ON	0.0
2955	2019-01-07	19:23:26	32	8412	2767.0	...	가동	ON	0.0
2957	2019-01-07	19:50:44	32	8412	2767.0	...	가동	ON	0.0
2959	2019-01-07	19:52:17	32	8412	2767.0	...	가동	ON	0.0

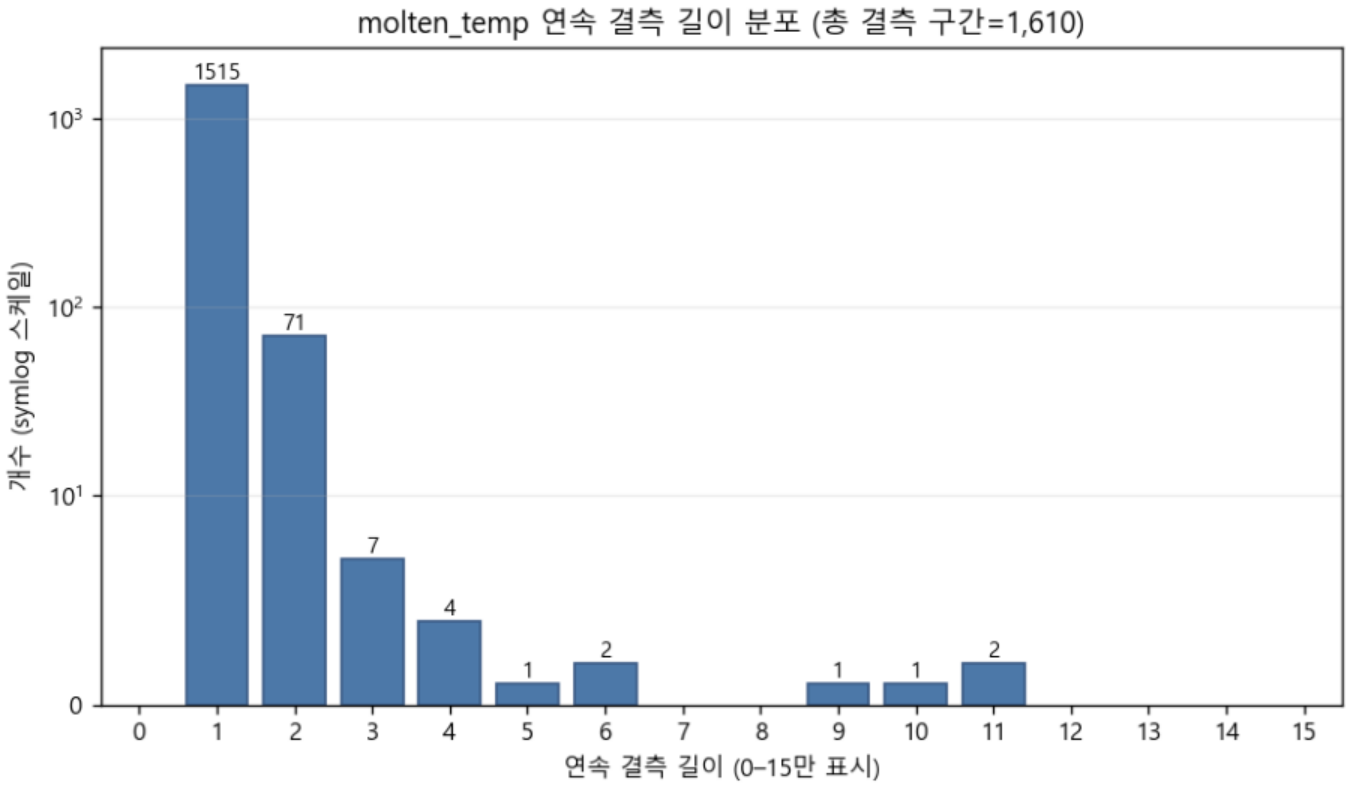


결측치 처리

결측치 대치 방법 설명

01 molten_temp 열

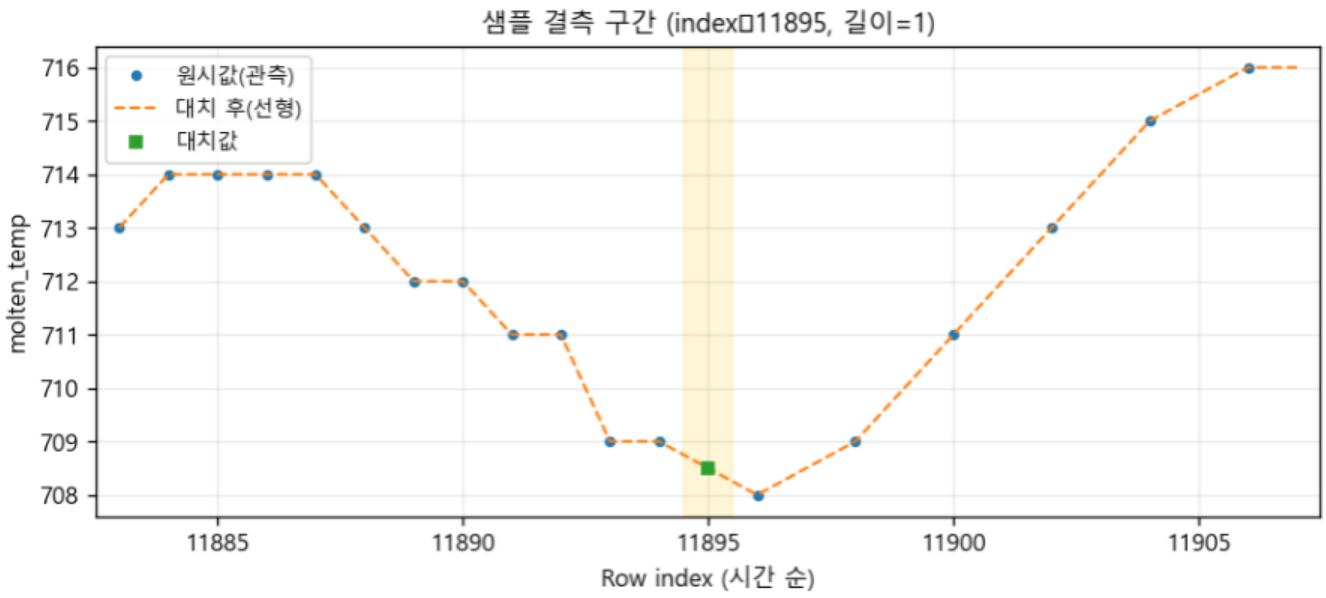
- molten_temp의 연속 결측 구간 길이 분포를 확인 → 1~6개 정도에 몰려 있음
  ⇒ 결측 구간이 길지 않으므로 선형 보간이 합리적임
- train 데이터에서는 결측이 있는 경우의 이전 행과 다음 행의 molten_temp 값들의 평균으로 대치
- test 데이터에서는 다음 행의 데이터를 알 수 없음 → 직전 행의 molten_temp 값으로 대치

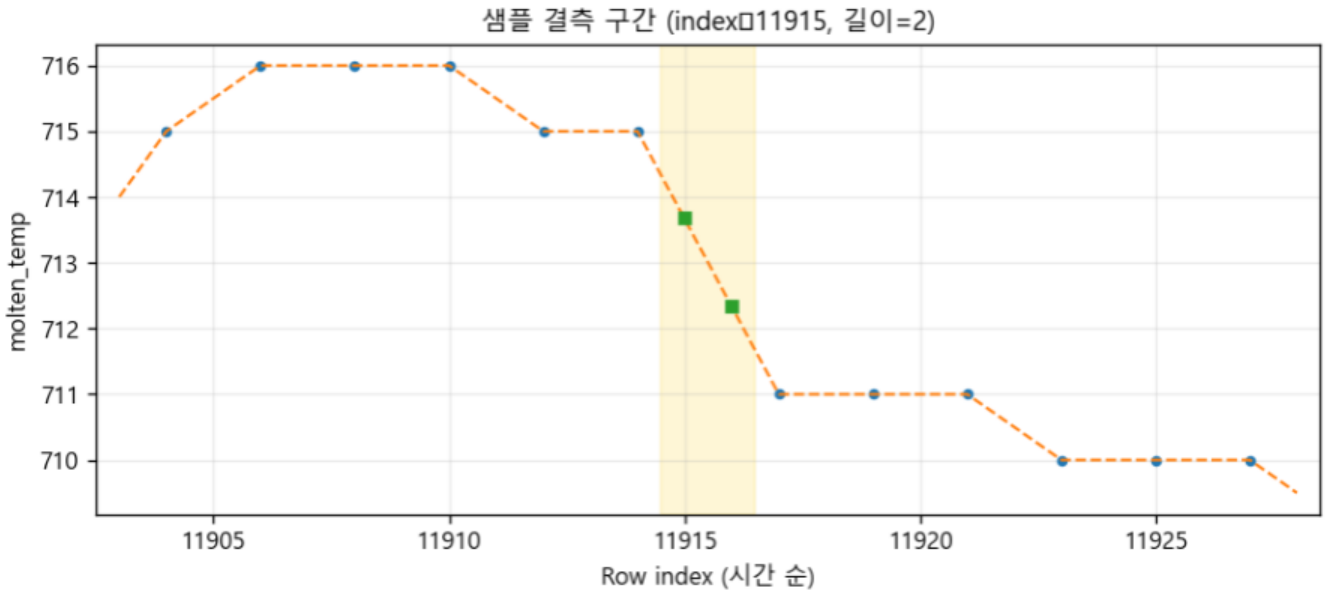


결측치 처리

결측치 대치 방법 설명

- 선형 보간으로 결측치를 대치한 이후의 molten_temp 그래프를 통해 대치법의 합리성 확인 가능





CONCLUSION

결론

데이터 전처리 내역 정리 & 전처리 결과 요약

데이터 전처리 내역

☑ 단일 칼럼 제거: heating_furnace 열, molten_volume 열,
upper/lower_mold_temp3 열, registration_time 열

☑ 행 제거: emergency_stop이 결측인 행, count가 중복인 행

☑ 결측치 처리: molten_temp 열 결측치를 선형 보간으로 대치

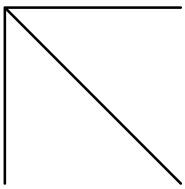
전처리 결과 요약

☑ 위의 단일 칼럼 제거 케이스 5개 포함 총 10개 칼럼 제거
(id, name, line, mold_name 등 구분용 변수 제거)

☑ 전체 행 중 위의 emergency_stop이 결측인 행, count가 중복인 행 등 총 1318개 행 제거

☑ 모델 학습에 사용한 train_df 크기: (72295, 21)
→ 전체 데이터 73612개 중 72295개만 사용 (98.2% 사용, 1.8% 제거)

2025.10.02



이상 보고를 마칩니다.

감사합니다.

3조 좋은데요