전처리Flow 및 근거 정리

1.가용변수 선택

-모델에 가용한 feature:

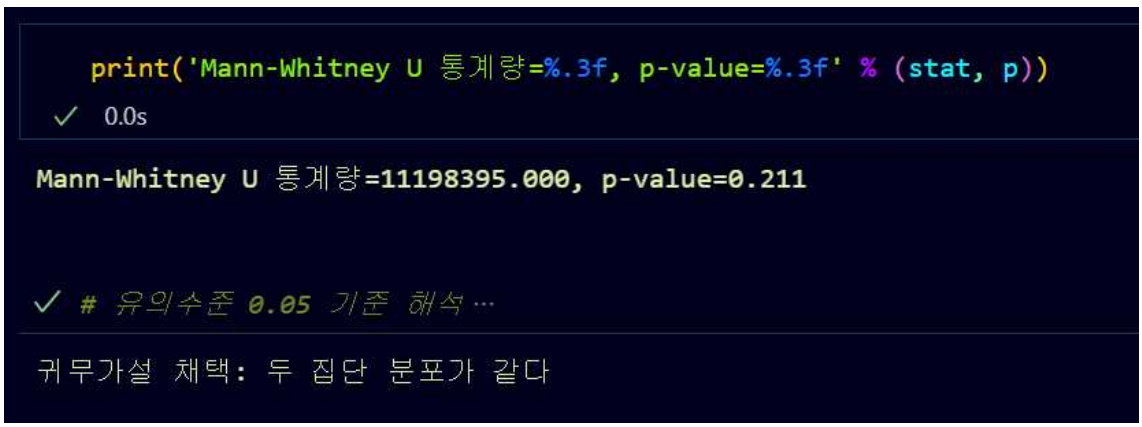
생산순번(count), 금형코드(mold_code), 가동여부(working), 시험샷여부(tryshot_signal)
생산 사이클 타임(production_CycleTime),
설비가동 사이클 타임(facility_operation_CycleTime),
용탕량(molten_volume), 용탕온도(molten_temp),
전자교반 가동시간(EMS_operation_time), 슬리브 온도(sleeve_temperature),
주조압력(cast_pressure), 비스켓 두께(biscuit_thickness), 저속구간속도(low_section_speed
고속구간속도(high_section_speed), 형체력(physical_strength),
상금형온도1~2(upper_mold_temp1~2), 하금형온도(lower_mold_temp1~2)
냉각수온도(Coolant_temperature), 양/불 판정(passorfail)

기초 시각화 EDA 및 반고상 주조 공정 과정의 이론적 근거를 통해 다음 변수들 선정
주요 영향 변수 (반고상 주조 관점)
- **molten_temp (용탕 온도)**
  → 단순 액상 온도 관리보다 *고상 분율(Solid fraction)* 제어가 핵심
  → 너무 높은 온도 → 고상 분율↓, 슬러리 유동성 과도 → 기공, 수축공
  → 너무 낮은 온도 → 고상 분율↑, 충전 불량·냉결(cold shut).
- **EMS_operation_time (전자교반 가동 시간)
  → 반고상 공정의 핵심. EMS 시간과 강도가 *고상 입자 크기·분포*에 직접 영향.
  → 불균일한 EMS → 국부 기공·미세조직 불균일 발생
- **facility_operation_CycleTime / production_CycleTime**
  → 반고상은 응고속도 민감도가 높음.
  → 사이클이 길면 입자 성장, 짧으면 충전 불완전.
- **low_section_speed / high_section_speed**
  → 반고상 슬러리는 점성이 높아 유동 제어가 더욱 중요.
  → 저속 충전 단계에서 *슬러리 응집 방지*, 고속 단계에서 *완전 충전* 필요.
- **cast_pressure (주조 압력)**
  → 액상 대비 슬러리 점성이 커서 *충분한 보압 유지*가 불량 억제 핵심.
- **biscuit_thickness (비스켓 두께)**
  → 보압 전달 통로 역할. 슬러리 충전에서는 더욱 중요.
- **mold temperature (upper/lower)**
  → 금형 온도 불균일 시 → 반고상 특유의 *고상 분율 불균일 응고* → 편석, 균열.
- **Coolant_temperature (냉각수 온도)**
  → 냉각 조건이 조직 미세화, 수축 결함 억제에 직결.
- **sleeve_temperature (슬리브 온도)**
  → 슬러리 응고 시작 지점. 슬러리 특성상 슬리브 응고 → 바로 불량.
- **physical_strength (형체력)**
  → 반고상은 충진 압력이 높기 때문에 금형 밀착 불량(벌어짐, flash) 가능성이 더 크다.

*tryshot_signal, count, working 변수는 공정초기화 직후 초기 생산품 및 시험생산품, 생산 중 공정 정지를 판단하는 변수로 활용하기 위해 함께 가용

-제외한 feature:

제품 id
작업라인(line): '전자교반 3라인 2호기'로 Value 동일
제품명(name): 'TM carrier RH' 로 Value 동일
금형명(mold_name): 'TM Carrier RH-Semi-Solid DIE-06' 로 Value 동일
비상정지여부(emergency_stop): 'ON'으로 Value 동일
수집시간(time): 생산사이클 진행에 따라 2분간격으로 시간대 정보 기록, EDA에 활용
수집일자(date): 생산일자 정보, EDA에 활용
등록일시(registration_time): date, time과 동일
가열로구분(heating_furnance):
값 count: 'A'(16413), 'B'(16318), 'NaN'(40881)
'C'라는 가열로 설비가 따로 있을 것이라고 생각하고 'nan' 값을 C로 대체
-같은 가열로 간 몰드 코드 별 용탕온도 차이 분포 검정(Mann-whitney U)



-> 용탕의 온도가 균일하게 이뤄지고 있다는 가정(몰드 코드 간)&가열로 별 불량률 차이 X 및 다를 지라도 금형 코드와 전자교반 시간(물질 특성)에 따른 공정 변수 별 분포 차이가 두 입력 변수를 통해 충분히 구분될 것이라고 판단하여, 가열로 구분 변수 제외

실제 가열로 별 차이를 금형코드와 물질특성 변수로 최대한 반영

상금형온도3(upper_mold_temp3): '1449'라는 동일 값 64356개(전체 73612행)
하금형온도3(lower_mold_temp3): '1449'라는 동일 값 71650개(전체 73612행)
*평균적으로 금형온도 100~200사이의 값
-> 사실상 상수 특성을 띄기에 예측에 유의미한 영향을 줄 수 없는 변수로 판단하여 제외

*양/불을 예측하는 모델에는 가용하지 않았지만 EDA 및 전처리 근거 확보, 분석 파이프라인 설계 시 근거 확보를 위해 가용

2.데이터 정제

-중복 행 제거: time 시간대 정보만 다르고, 동일한 생산 제품 연속적으로 등장한 데이터 10개 행 제거

time	date	count	emerge	molten	facility_	produc	low_se	high_se	cast_pr	biscuit_	upper_	upper_	upper_	lower_t	lower_t	lower_t	sleeve_	physica	Coolan	EMS_o	passorf	mold_c
9:15:04	2019-03-05	34	ON	730	117	136	110	113	320	52	200	118	1449	122	198	1449	361	702	32	6	0	8412
9:15:47	2019-03-05	34	ON	730	117	136	110	113	320	52	200	118	1449	122	198	1449	361	702	32	6	0	8412

-데이터 타입 변경: 'mold_code', 'EMS_operation_time' (범주형 변환)

금형 코드 고유 범주값: 8412, 8573, 8600, 8722, 8917

전자교반 시간 고유 범주값: 3, 6, 23, 25

-> 해당 범주 유형마다 제품 생산과정에서 공정 변수별 작용이 다를 것이라 예상하여 범주화

-결측치 처리

***행 제거: 19327(id)**

***대치:**

시험샷여부 NaN 값: 'A' 라는 정상 생산품을 나타내는 범주 값으로 라벨링

용탕량 NaN 값: (-1)의 음수 값으로 처리

전체 행 중 절반 이상 결측: 34992행

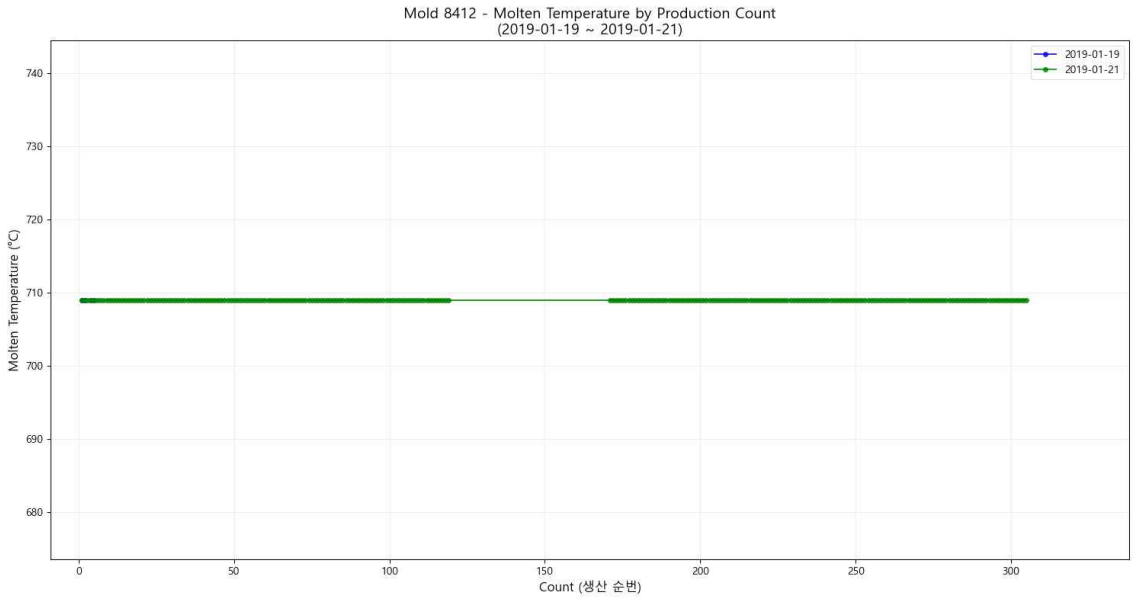
가열로 A, B 일 때 모두 결측값(가열로 A,B 용탕량 측정 안됨)

각 용탕마다 크기나 금속용액 소모 패턴이 다를 수 있기 때문에 가열로 C의 소모 경향으로 3만개 이상의 결측치를 채우는 것이 위험하다고 판단하여 (-1)의 임의 값 부여 후 모델 투입

용탕온도 NaN 값: 결측치 발생 직전 시점(1순번 이전 제품), 결측치 발생 직후 시점(1순번 이후 제품) 값인 선형보간 값 709로 채우기

용탕온도 결측치 발생 케이스는 금형코드 8412 제품 생산 중 1월 19일 01시(id:11895)~1월 21일 14시(id:14416)까지 결측 발생, 1월 23일 0:31:44(id: 16008), ~ 1월 24일 07:59:34(id: 17598) 까지 결측 발생

-> 연속적인 생산 시퀀스임을 고려하여 결측치 행들 앞,뒤 순번의 값으로 선형보간



<결측구간>

-이상치 처리

***행 제거: [42632, 35449, 6000, 11811, 17598, 46546, 35451]**

*센서 입력 오류 의심값 중 연속적으로 발생되는 값이 아닌 행들(일시적인 스파크 값으로 의심되는) 및 대부분 변수에서 결측치를 갖는 행 제거(추후 안정적인 모델링을 위해서)
ex) 구간속도, 온도, 형체력 등 6만 대 값을 갖는(원 스케일은 두자리~세자리수 단위)

***대치:**

생산 사이클타임 0 값: 설비 가동 사이클 타임으로 대체

일부 행들을 제외하고는 생산 사이클 타임과 설비 가동 사이클 타임 값이 오차범위 내에서 같은 경우의 패턴이 많아 해당 방식으로 대체

용탕 온도 0값:

0값을 갖는 행들의 경우 생산 사이클 중간 중간 산발적으로 발생하는 경향이 많아 0이 아닌 순번의 앞뒤 값으로 선형 보간

슬리브 온도 1449값:

1449값을 갖는 행들의 경우 8917 금형의 (3/06 13:03 ~ 3/07 08:25) 시간대 생산 사이클 중 산발적으로 등장한 케이스만 있어서 마찬가지로 생산 순번 앞뒤 값으로 선형 보간

냉각수 온도 1449값:

1449값을 갖는 행들의 경우 1/19일 00시부터 생산 사이클 중 9개의 연속적인 값으로 등장하는 케이스만 있기에 생산 순번 다음 값인 35로 대치

형체력 0값 중 1/02~1/03 7am 생산 제품 행:

해당 일자의 8412 금형 제품 생산 전체 사이클에서 형체력이 0의 값을 보이지만, 전부 양품이며 다른 변수 값들은 거의 대부분 정상 범위 내에 있기에 **1)0의 값은 입력 오류로 판단**
해당 일자 전체 사이클에서 통으로 0의 값을 가져 선형 보간할 참조 대상이 없기에,
2)0의 값은 8412 금형의 형체력 평균으로 대체

*해당 일자 제외 다른 일자에서 발생하는 0의 값의 경우, 모두 불량 판정을 대치 X

3.모델링 준비

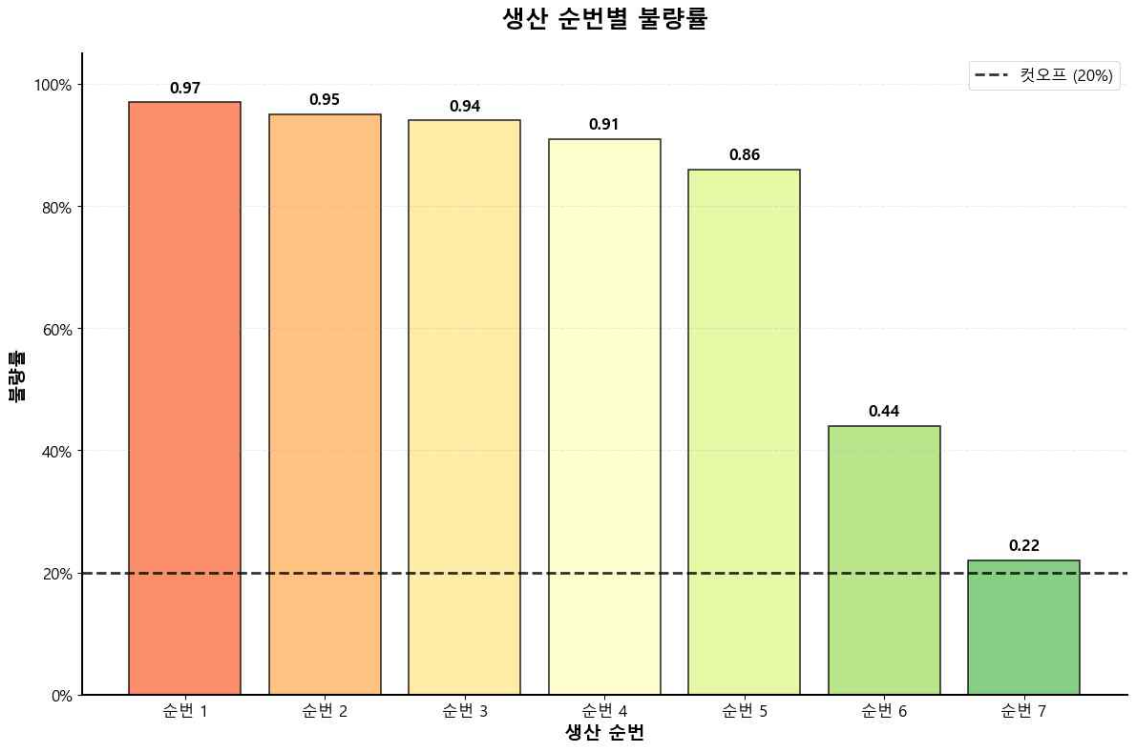
1)생산 사이클 정보

몰드코드 별 'Count' 순서 값이 생산 순번

***일자별 생산 사이클**

주간 생산 사이클: 약 오전 8시 ~ 약 오후 7시 (평균 2분 간격 생산)
야간 생산 사이클: 약 오후 8시 ~ 약 오전 7시 (평균 2분 간격 생산)
공정 초기화 직후 8am, 8pm 제품의 불량률 0.2 이상
*불량률 100% 범주: 시험샷여부 = D, 가동여부 = 정지

높은 확률로 불량패턴(생산 사이클 초기-공정 불안정)



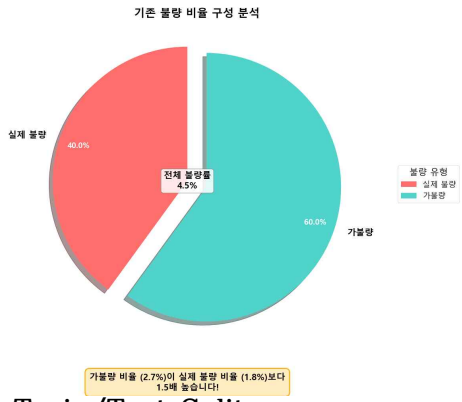
2)실제 불량 샘플 라벨링

제품 생산 과정에서 공정 초기화 직후 재생산 시 초기 생산품, 시험샷의 경우 공정 조건 불안정화로 인한 거의 대부분 불량 제품이 생산됨. 이러한 경향에 따라 정상 생산 과정 중 발생하는 불량에 대한 예측을 중요시하는 모델을 개발하기 위해 불량 라벨을 1)실제 불량, 2)가불량으로 나누어 판단

-가불량 판단 기준:

1) Count를 기준으로 불량률 0.2 이상인 Count 값을 갖는 불량 샘플에 대해 가불량 판정
2) Tryshot_signal이 D인 샘플에 대해 가불량 판정
(실제 불량은 가불량 판정 샘플 제외 불량 샘플)

->실제 불량만 1로 라벨링 한 변수 '**realfail**' 생성
->양품, 가불량, 실제불량 샘플에 대해서 라벨링한 변수 '**check**' 생성
양품: 0, 가불량: 1, 실제 불량: 2



-Train/Test Split

8:2 비율로 금형코드, 양/실제불/가불 로 라벨링된 check 변수 비율에 맞게 층화 샘플링

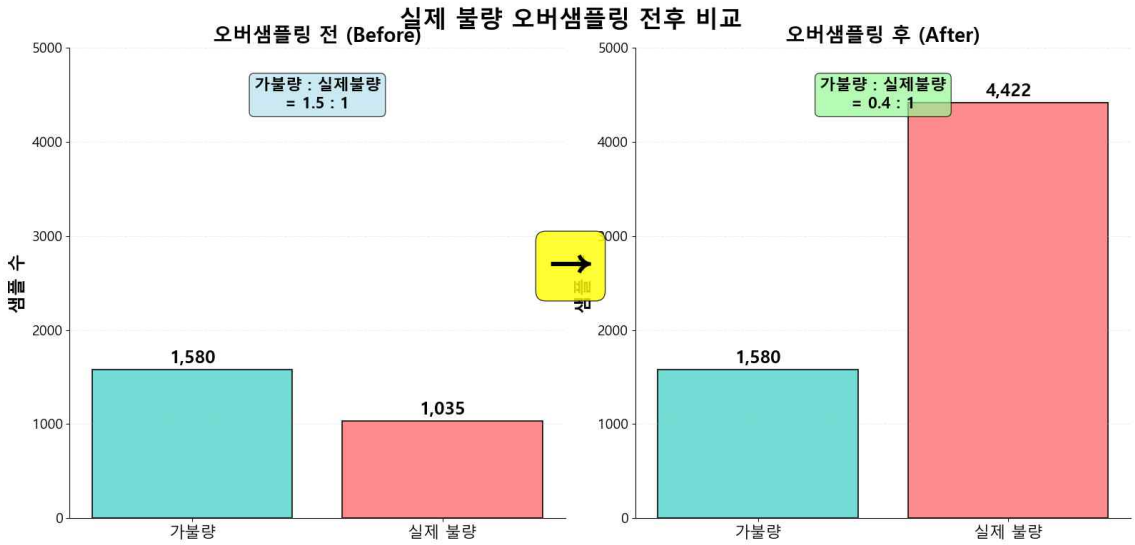
-실제 불량 샘플에 대해 사전 오버샘플링(Train data)

사전에 생성하였던 실제 불량이 라벨링 된 변수를 활용해
train 데이터에서 실제 불량에 해당하는 샘플에 대해서
몰드코드별 비율을 유지한 채 4배 오버 샘플링(Smote 활용, 범주형은 voting 방식 채우기)

양성비 변화(그룹별):

mold_code	before	after
8412	0.007585	0.030397
8917	0.018138	0.072589
8722	0.019836	0.079369
8573	0.024568	0.098279
8600	0.038127	0.152636

-오버샘플링 이후 가불량 대비 실제 불량률 비: 약 2.5배


실제 불량 오버샘플링 전후 비교
오버샘플링 전 (Before)
가불량 : 실제불량
= 1.5 : 1
샘플 수
5000
4000
3000
2000
1000
0
1,580
1,035
가불량
실제 불량
오버샘플링 후 (After)
가불량 : 실제불량
= 0.4 : 1
4,422
1,580
가불량
실제 불량