4주차 실습 프로젝트

다이캐스팅 제조 데이터 불량예측

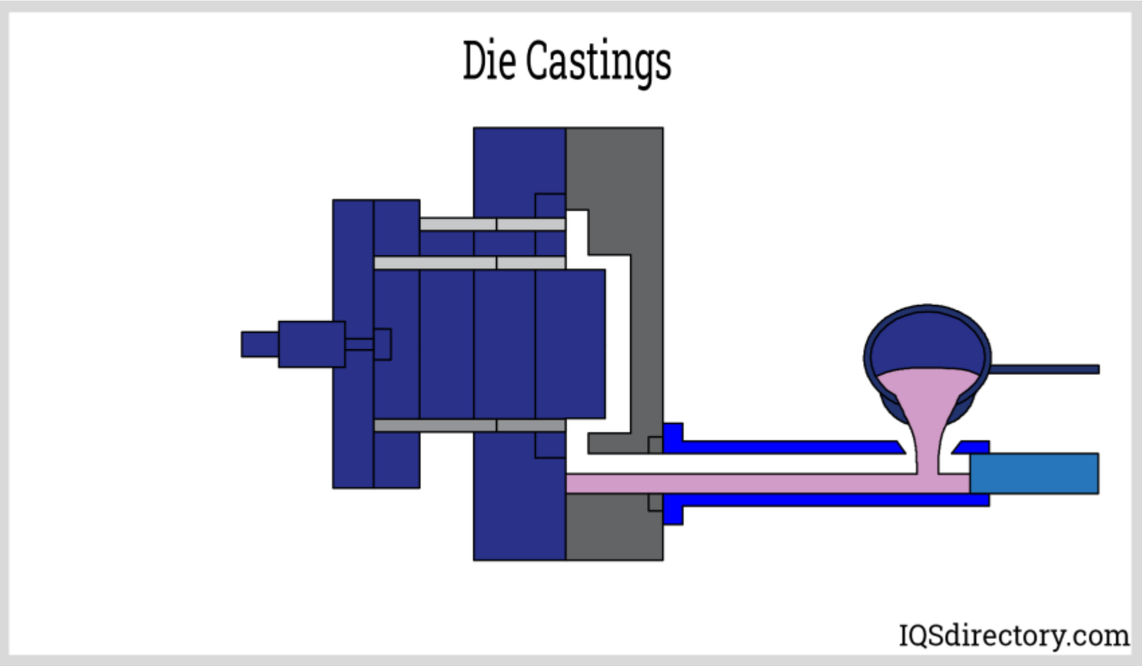
LS Big Data School Group 4
윤주영, 김서정, 박성민, 서성호, 이채원, 임유빈

CONTENTS

팀 소개 및 역할분담

훈련생	주역할	담당업무
윤주영	**팀장**	데이터 시각화 및 모델링
김서정	팀원	PPT 제작 및 내용 정리
박성민	팀원	데이터 전처리 및 EDA
서성호	팀원	결측치 및 이상치 분석
이채원	팀원	데이터 시각화 및 분석
임유빈	팀원	자료수집 및 모델링

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

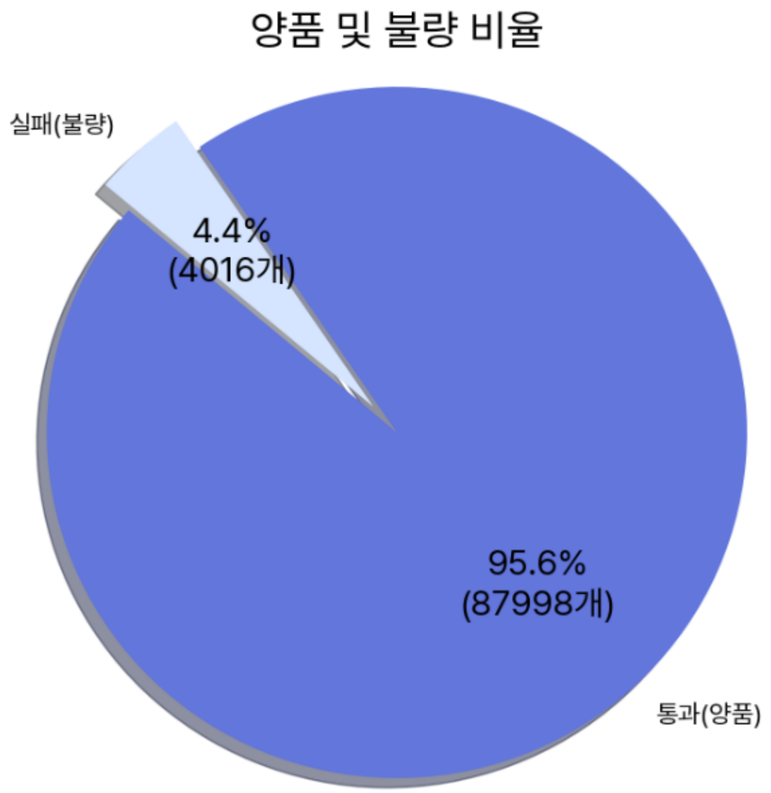
구분	명칭
독립변수	용탕온도(molten_temp)
	제품 생산 사이클 시간 (production_CycleTime)
	저속구간속도(low_section_speed)
	고속구간속도(high_section_speed)
	주조압력(cast_pressure)
	비스켓 두께(biscuit_thickness)
	상금형온도1(upper_mold_temp1)
	상금형온도2(upper_mold_temp2)
	상금형온도3(upper_mold_temp3)
	하금형온도1(lower_mold_temp1)
	하금형온도2(lower_mold_temp2)
	하금형온도3(lower_mold_temp3)
	슬리브온도(sleeve_temperature)
	형체력(physical_strength)
	냉각수온도(Coolant_temperature)
종속변수	양품불량판정(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
  - 범주형 : 작업라인, 제품명, 금형명, 수집시간, 수집일시, 가동여부, 비상정지, 등록일시, 사탕신호, 가열로
  - 정수형 : 일자별 제품 생산 번호, 설비 작동 사이클시간, 제품생산 사이클 시간, 전자교반 가동시간, 금형코드
  - 실수형 : 용탕온도, 저속구간속도, 고속구간속도, 용탕량, 주조압력, 비스켓 두께, 상금형온도1-3, 하금형온도1-3, 슬리브온도, 형체력, 냉각수 온도, 양품불량판정

데이터 전처리 및 시각화

1. 양품 및 불량 개수 확인



불량률이 극도로 낮은 경우 → 학습 및 테스트 데이터셋이
비대칭적으로 구성될 수 있음을 확인!

2. 숫자형 변수만 사용하기

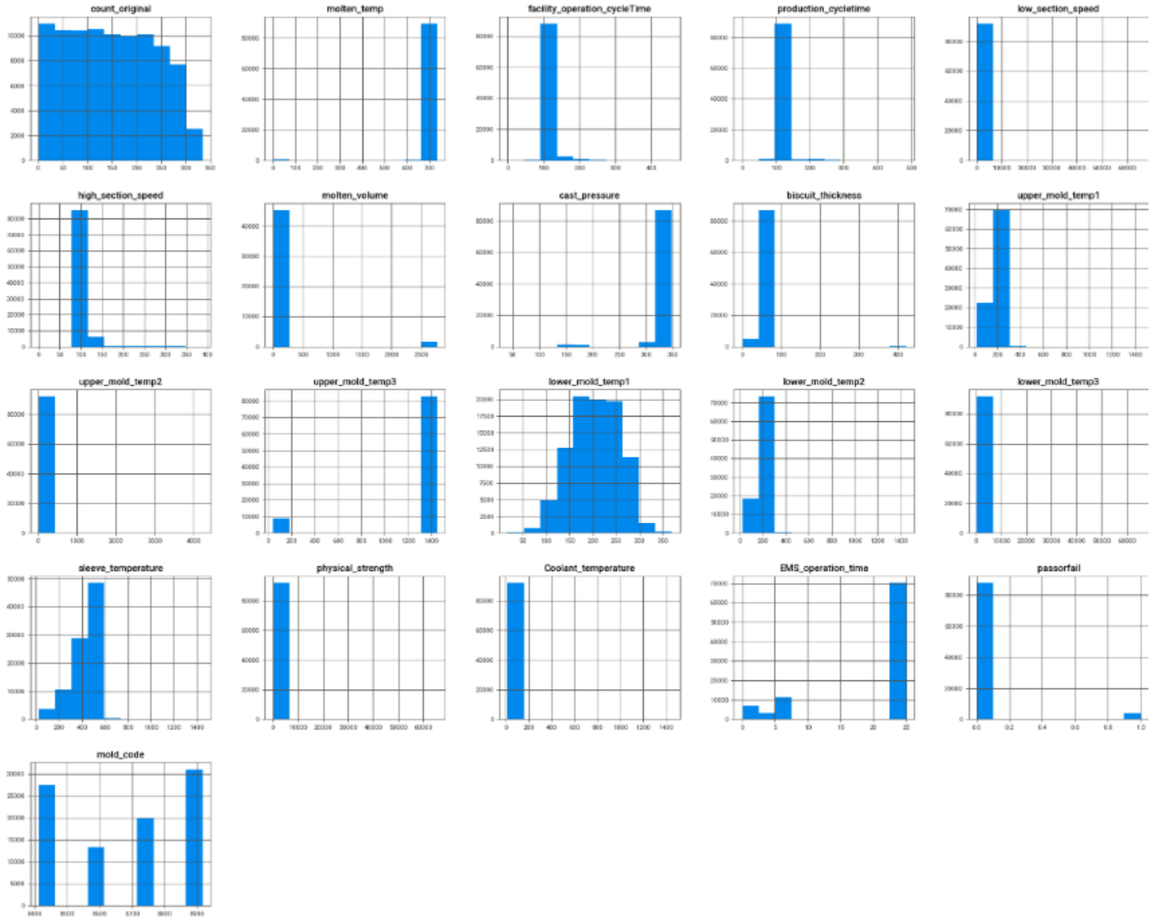
숫자형이 아닌 변수로는 학습할 수 없기 때문에 일반적으로 문자열을 숫자로 변환하거나 숫자만 사용한다. 이 데이터셋에는 숫자형 데이터만 필요하므로 'object' 타입(문자열)이 아닌 변수들로 데이터를 재구성한다.

범주형 변수로는 작업라인, 제품명, 금형명, 수집시간, 수집일시, 가동여부, 비상정지, 등록일시, 사탕신호, 가열로가 포함된다.

object 변수명	
작업라인	line
제품명	name
금형명	mold_name
수집시간	time
수집일시	date
가동여부	working
비상정지	emergency_stop
등록일시	registration_time
사탕신호	tryshot_signal
가열로	heating_furnance

데이터 탐색

2. 히스토그램



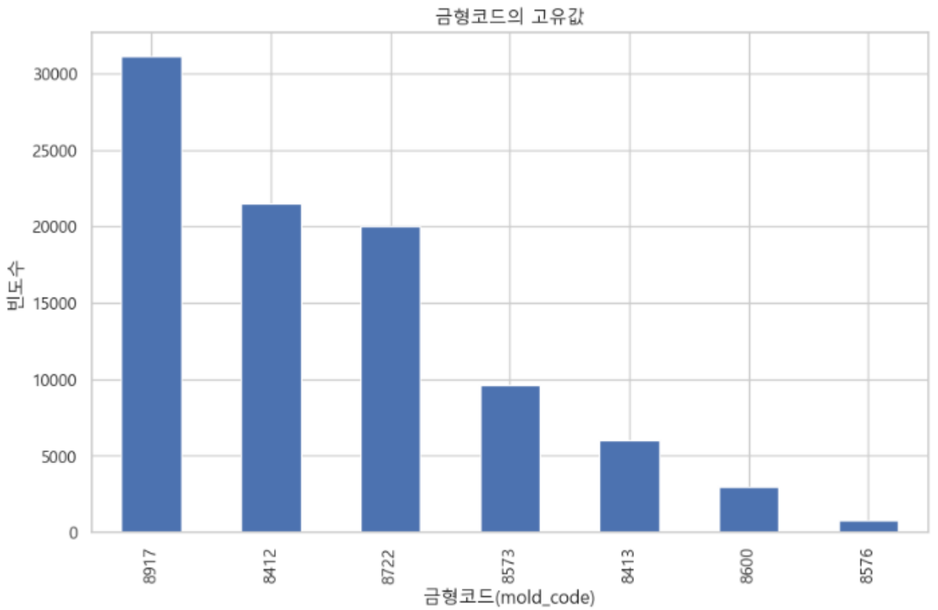
히스토그램으로 데이터 분포 확인!

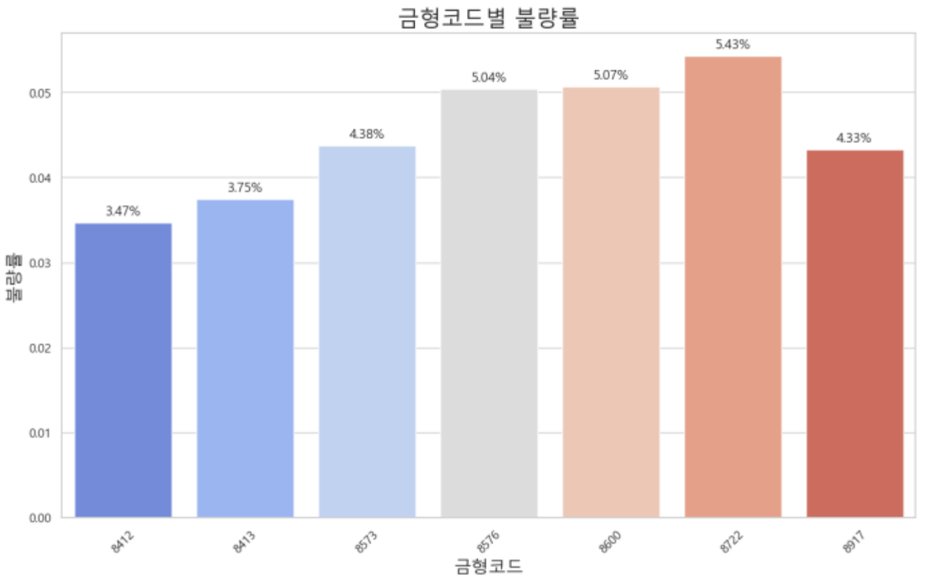
히스토그램 시각화를 사용하면, 정규분포에서 크게 벗어난 값들을 확인하여 이상치 존재 여부를 확인할 수 있다.
그러나, 공정 관리를 위한 특정 변수들은 **정규분포를 따르지 않는 경우**가 많다. 즉, **공정 최적화**를 위해 **특정값에 고정된 변수**들이 존재하기 때문에 히스토그램만으로는 이상치를 판단하기 어렵다고 생각하였다.

그렇다면, 공정 변수 별로 정규분포를 따르는지, 특정 값으로 고정된 값인지를 판단해서 이상치 여부를 신중하게 판단해야겠어!

데이터 탐색

3. 변수값 확인





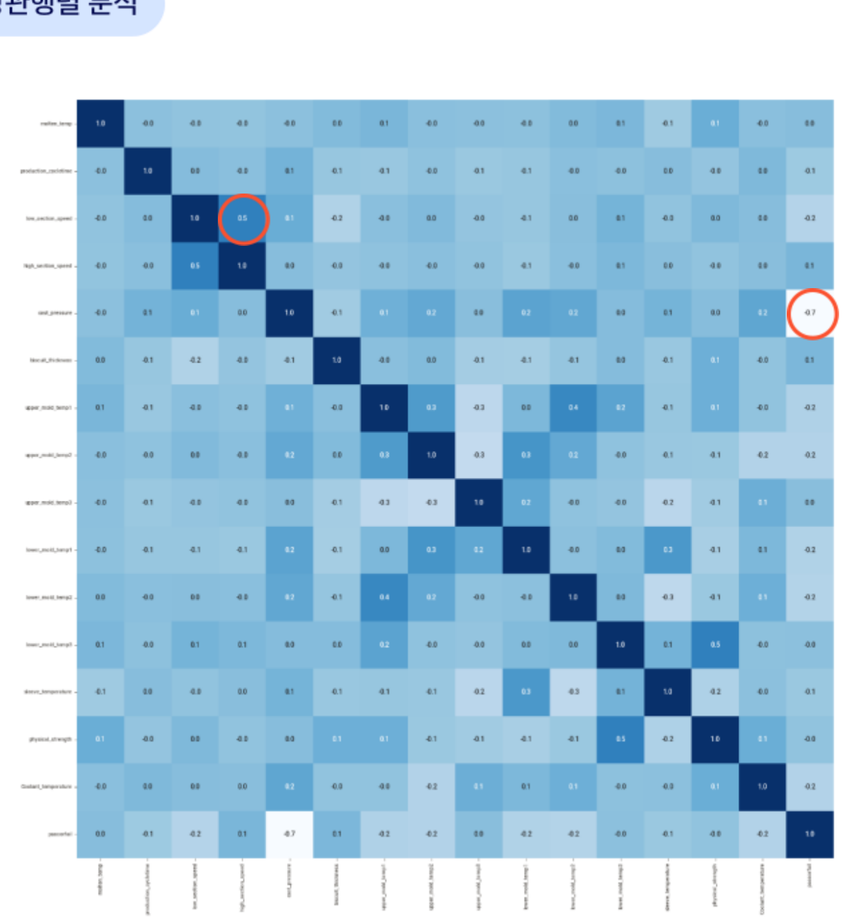
금형코드(mold_code)는 정수형 변수로, 7개의 고유값을 가진다. 데이터 빈도 수는 8917, 8412, 8722, 8573, 8413, 8600, 8576 순이다.

→ 빈도수가 높은 금형코드의 제품에 더 많은 자원을 투자하여 품질관리를 하는 것이 선택과 집중 측면에서 효율적일 수 있다.

금형코드 별 불량률을 시각화한 결과, 불량률이 가장 높게 나온 코드는 8772이며, 5.43%의 불량률이 측정된다. 불량률이 가장 낮게 나온 코드는 8412이며, 3.47%의 불량률이 측정된다. 전체 금형코드별 불량률은 3%에서 5% 사이에 분포하며, **특정 코드에서 불량이 많이 발생하지는 않는 것으로 확인**되었다.

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!

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 활용하였을 때, 모델의 성능이 더 좋아지는지를 판단하고자 한다.

☑**이상치의 상·하한 설정 범위를 조정**해서 최적의 성능을 나타내는 범위를 결정하는 방법에 대해 알아보고자 한다.

2. 모델 성능 향상

모델 성능 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

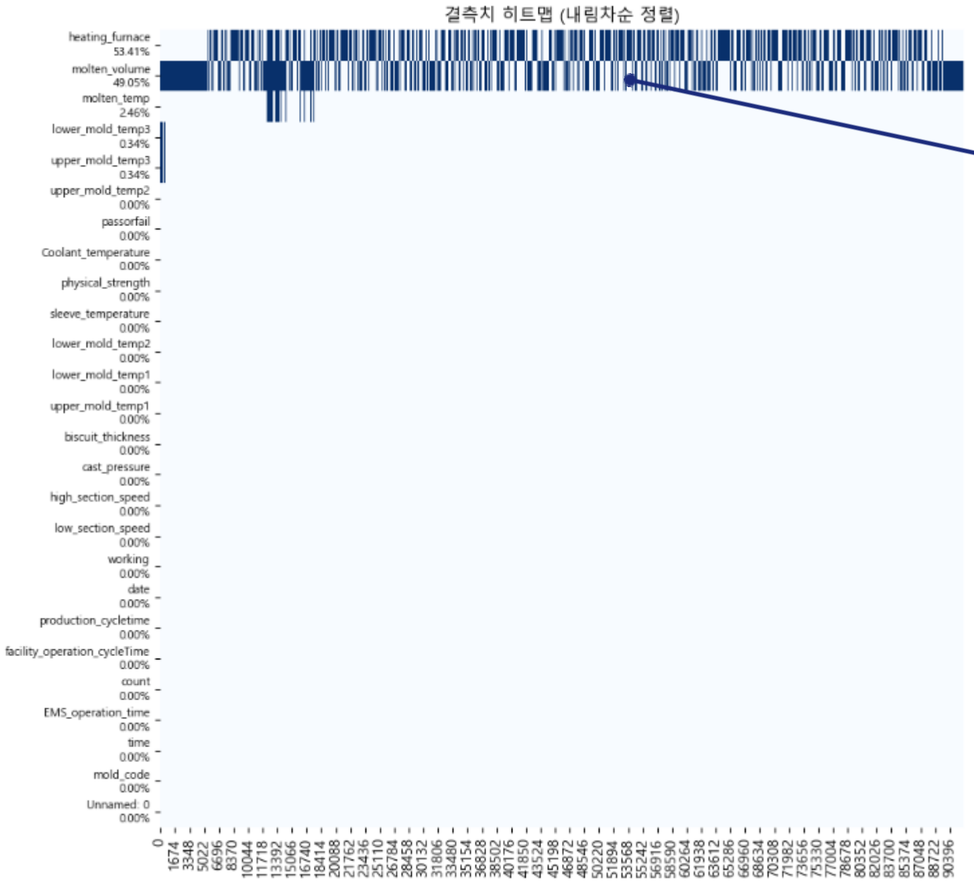
<수정 방향>

☑**불량 판단 모델**에 맞는 적절한 학습모델 선택을 통해 **예측 정확도**를 높일 수 있는 방안에 대해 탐색해보고자 한다.

☑ 데이터셋에 존재하는 불균형을 해소하기 위해 **SMOTE 기법**을 사용해보고자 한다.

결측치 처리 방법

1. 결측치 처리에 따른 모델 성능



moltem_volume

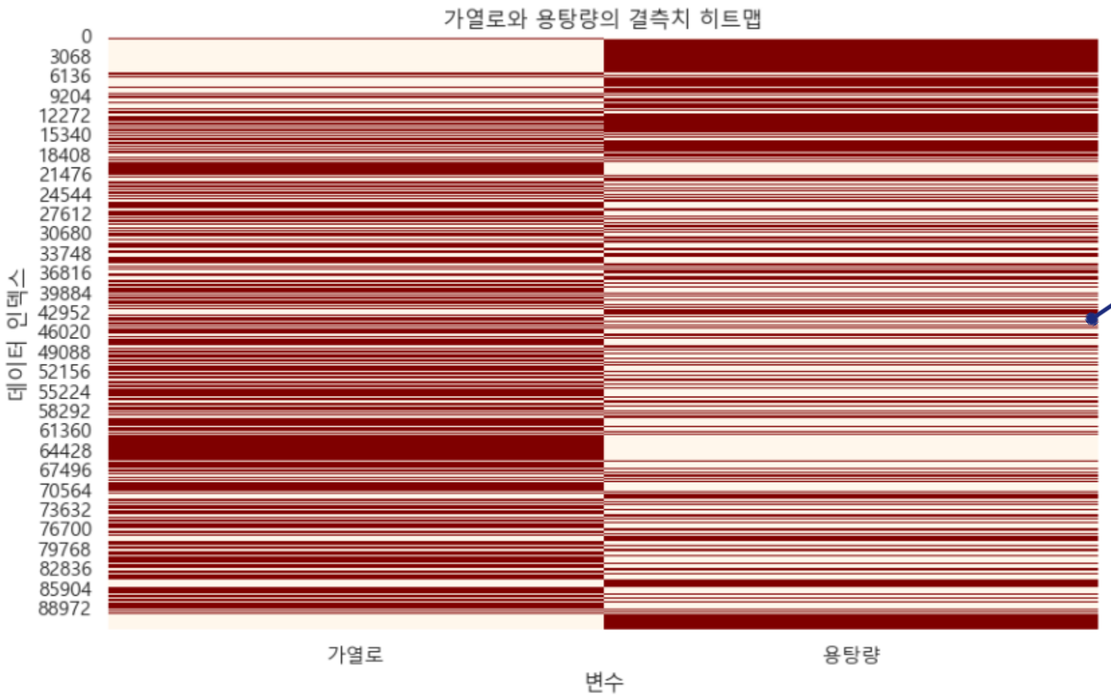
용탕량(molten_volume)은 전체 데이터 개수 중 50%가 결측값을 가진다.

▼

용탕량(molten_volume)이 다른 변수에 비해 결측값이 많은 이유는 무엇일까?

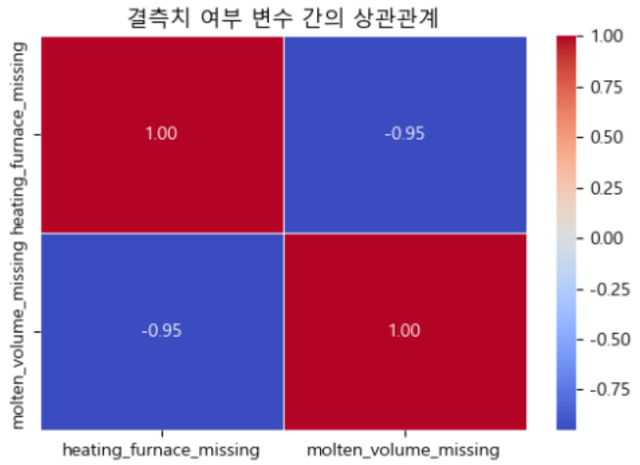
결측치 처리 방법

1. 결측치 처리에 따른 모델 성능



molten_volume VS heating furnance

가열로의 결측치가 있는 부분과 용탕량의 결측치가 있는 부분이 서로 상충된다.



결측치 처리 방법

1. 결측치 처리에 따른 모델 성능

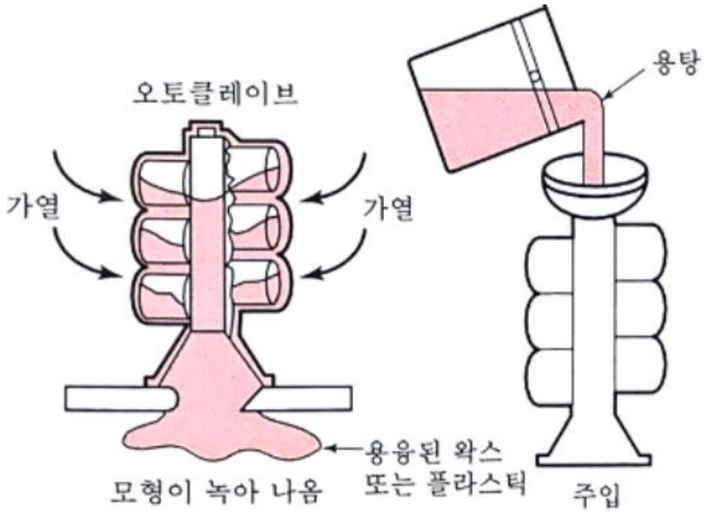
용탕량과 가열로 결측치 결론

용탕량(molten_volume)과 가열로(heating furnance)의 결측치를 히트맵을 통화 시각화하여 **두 결측치가 상호 상충관계**에 있다는 것을 알게 되었다.
두 변수의 결측치는 일부 겹치는 데이터를 제외하고 상관관계가 -0.95에 달할 만큼 높은 상관관계를 가지고 있다.

<해석>

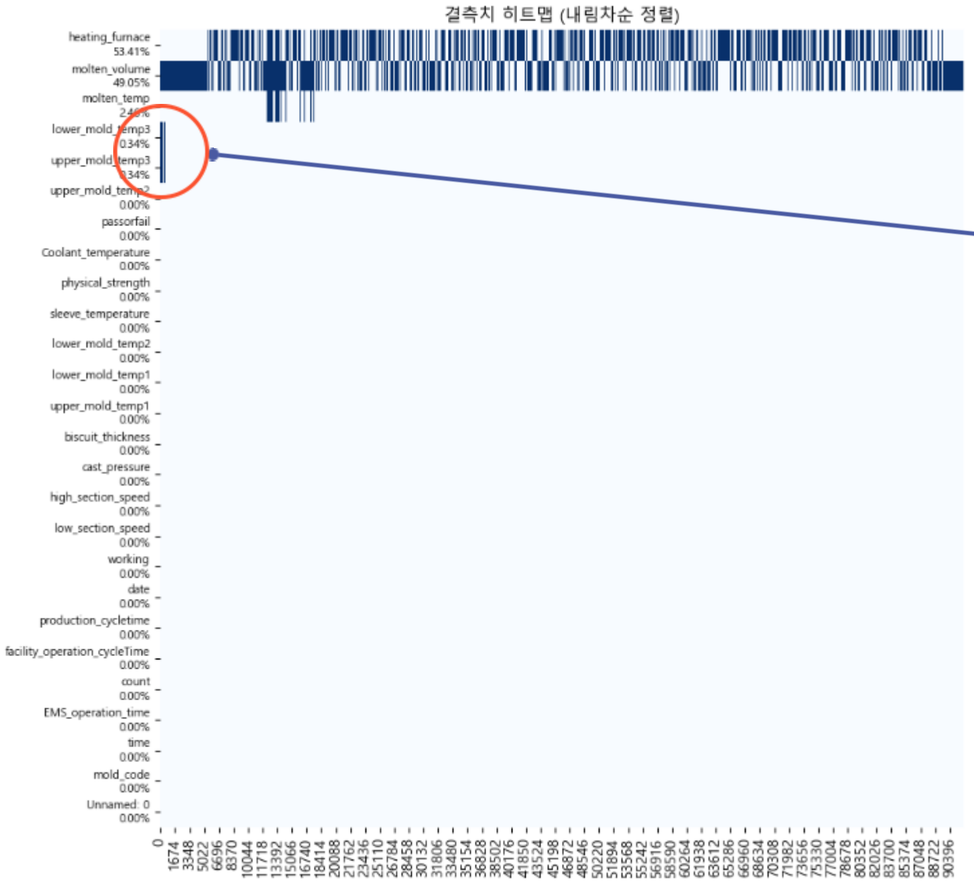
1) 주조 공정에서 금속을 가열로에 녹인 다음 용탕로에 담아 용탕량을 측정하는 공정 과정을 거치기 때문에 해당 변수들의 결측값이 상충하는 관계를 가진다.

2)이는 결측치가 특정 오류로 인해 발생한 것이 아니라 **공정 자체에서 발생한 것**임을 의미합니다. **따라서 결측치를 drop하고 모델링을 진행하기로 결정하였다.**



결측치 처리 방법

1. 결측치 처리에 따른 모델 성능



lower_mold_temp3 VS upper_mold_temp3

상금형온도3과 하금형온도3의 결측치 개수는 313개로 동일하며 결측치 분포가 유사한 것으로 파악된다.

▼

두 변수의 결측치 개수와 분포가 동일한 이유는 무엇이며, 공정에서 어떤 관계를 가지고 있을까?

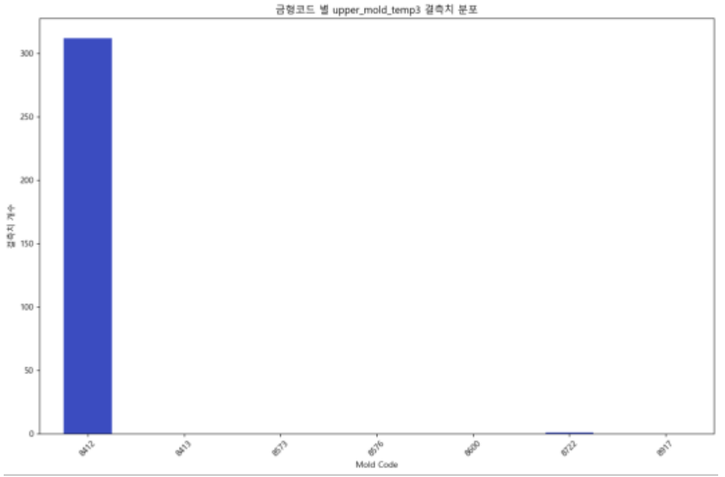
결측치 처리 방법

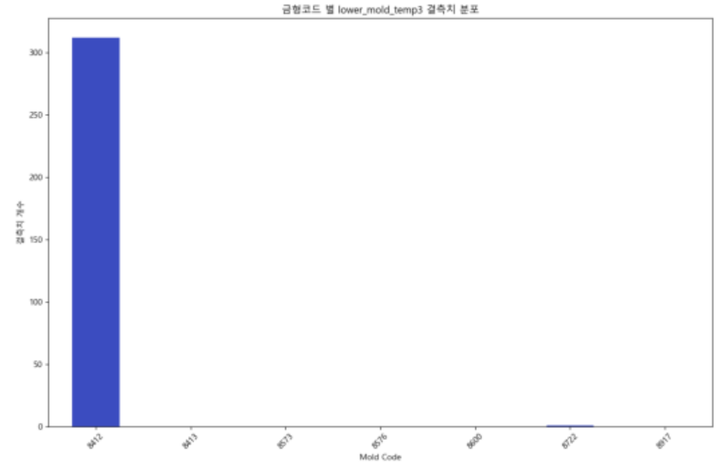
1. 결측치 처리에 따른 모델 성능

상금형온도3와 하금형온도3

상금형온도3과 하금형온도3의 결측치는 전체 데이터의 0.3%로 통계적으로 유의미한 차이를 만들지 않는다. 따라서, 보간을 하기 보다 drop하는 것이 더 적절한 방법이라고 판단하였다.

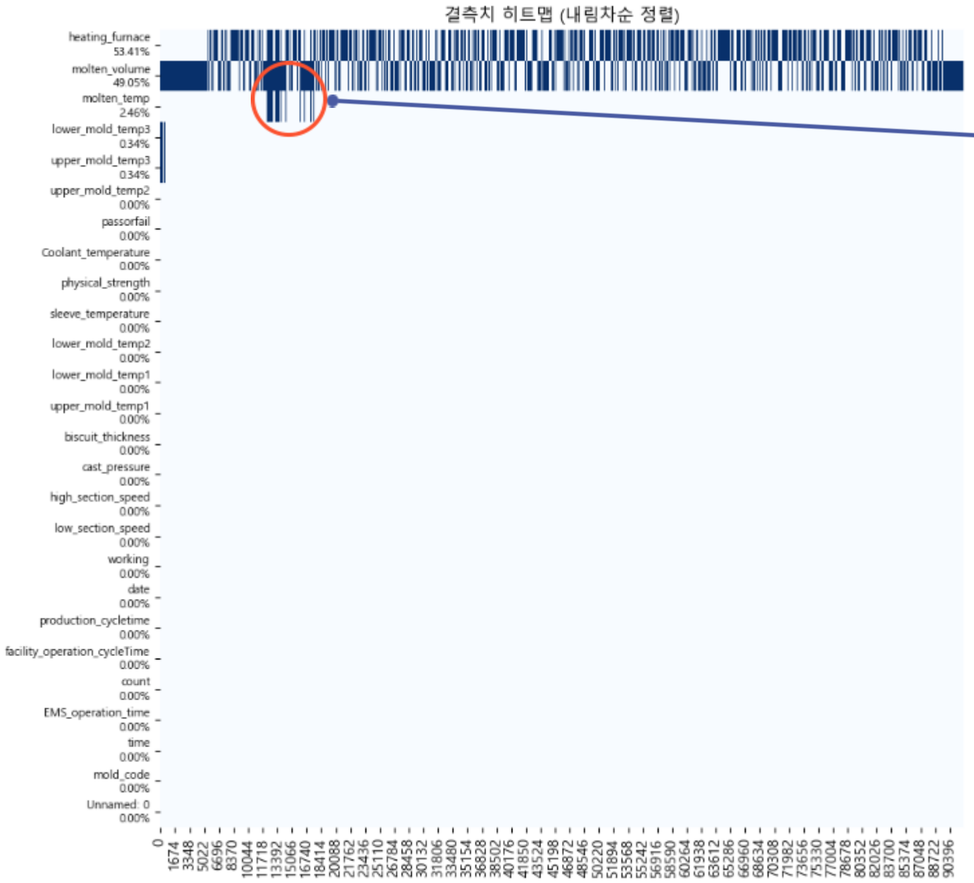
```
'molten_temp'결측치에서 불량율 비율: 0.028306059265811 59
```





결측치 처리 방법

1. 결측치 처리에 따른 모델 성능



molten_temp

molten_temp는 2261개의 결측치를 가진다. 측정오류인 건지, 특정 상황에서 결측이 존재하는건지 파악해야 한다.

▼

molten_temp의 결측치를 임의로 채우는게 좋을지, 그대로 두는게 좋을까?

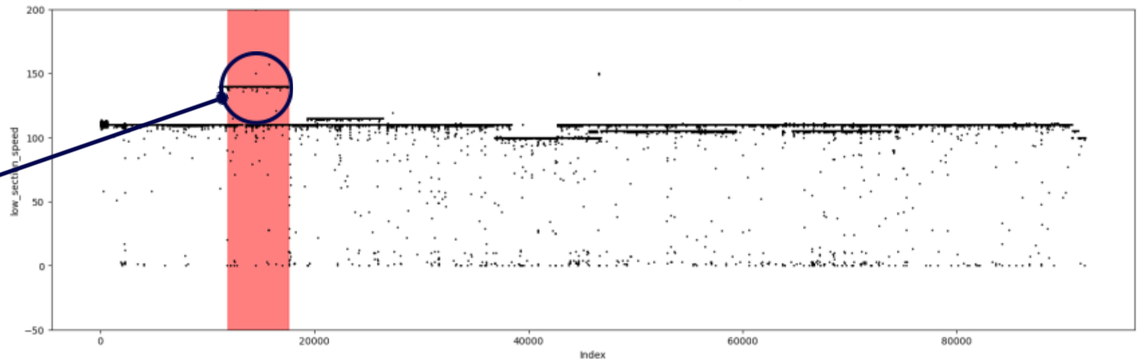
결측치 처리 방법

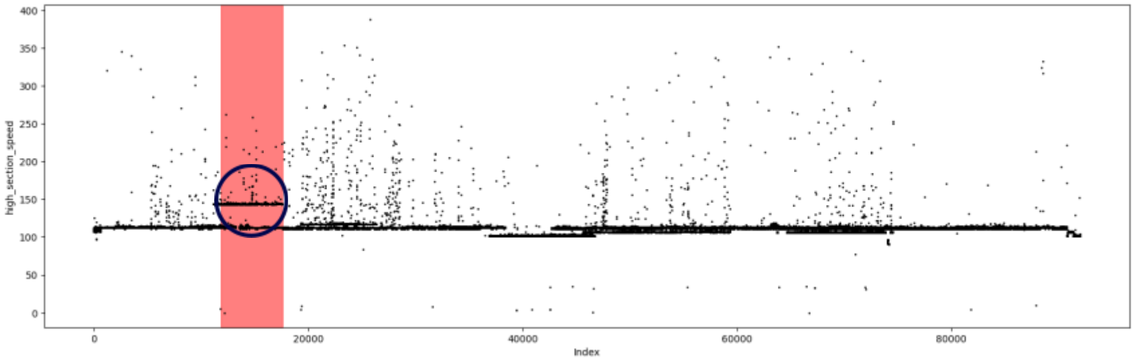
1. 결측치 처리에 따른 모델 성능

- production_cycletime : 주조 사이클이 완료되는 데 걸리는 시간
- low_section_speed : 금속이 주입될 때 초기 단계의 속도
- high_section_speed : 금속이 주입되는 후반 단계의 속도
- sleeve_temperature : 슬리브(주조 금형의 특정 부분)의 온도

molten_temp

분홍색으로 표시된 영역은 molten_temp의 결측치가 존재하는 범위이다. 이 영역의 표시를 유지한 채로 인덱스 별 변수의 값을 시각화한 그래프를 보았을 때, 총 4개의 열이 결측치 범위에서 독특한 값을 가지고 있는 것을 확인하였다.





결측치를 임의의 값으로 채워넣는다면, 4개의 변수에 영향을 미칠 가능성이 존재한다. 따라서 결측행을 제거한다.

이상치 처리 방법

2. 이상치 처리에 따른 모델 성능

이상치를 처리해야 하는 이유?

이상치는 정상적인 공정과정에서 수집되는 데이터의 범위를 넘어서는 것으로
이상치가 존재할 경우 모델의 정확도가 떨어지게 된다.

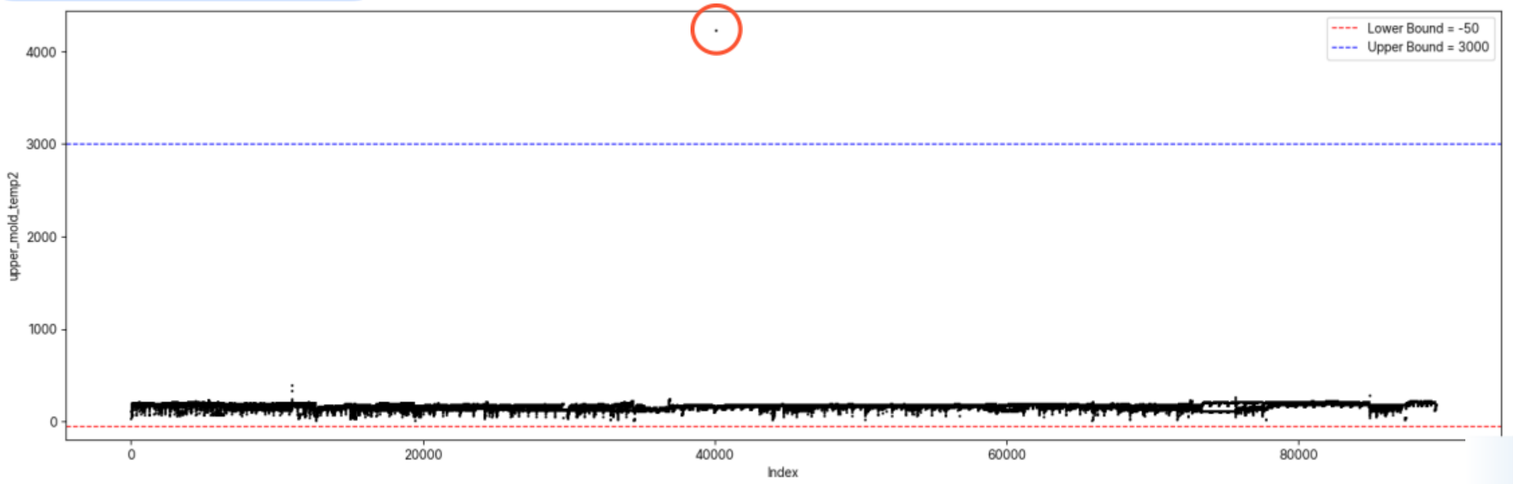
1. 기존 보고서에서는 변수들에 존재하는 이상치를 상하한 0.1%에 해당하는 값으로 제거했다.
2. 하지만 IQR 적용 시, 많은 양의 데이터를 제거하게 되어 데이터 부족 및 불균형 현상이 나타날 수 있다.

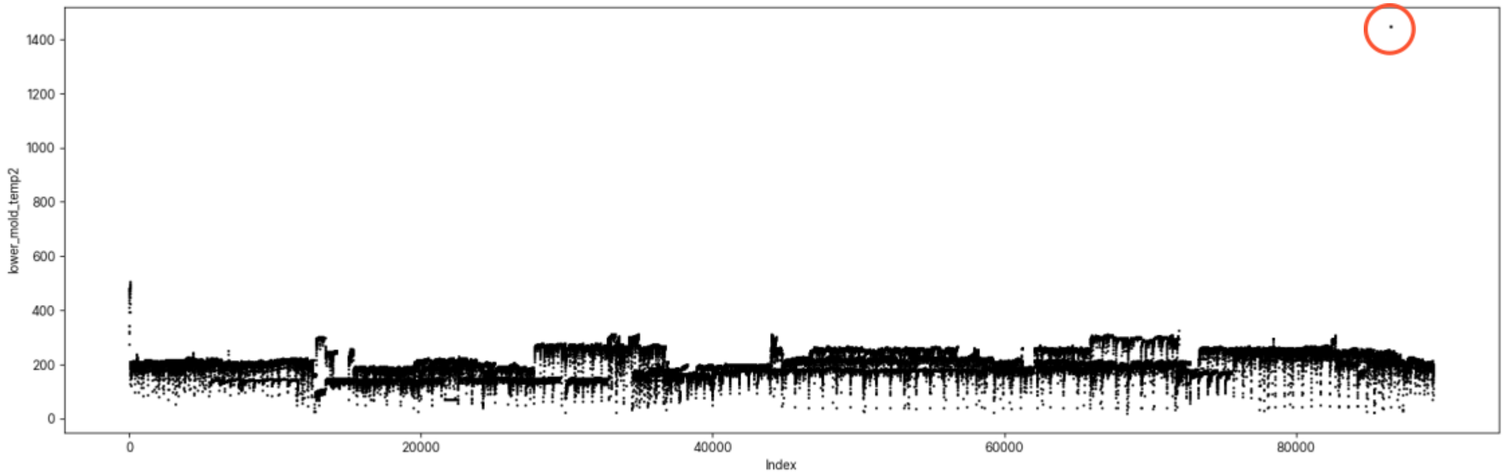
**31개의 열들 중에서 'object' 타입을 전부 제거한 후,
나머지 열들을 전부 plot하여 시각적으로 존재하는 이상치를 각각 확인해본다!**

- 이상치 제거를 통해 모델의 정확도를 향상시킨다.
- 모델 성능 향상을 통해 불량품 발생을 줄인다.

이상치 처리 방법

2. 이상치 처리에 따른 모델 성능

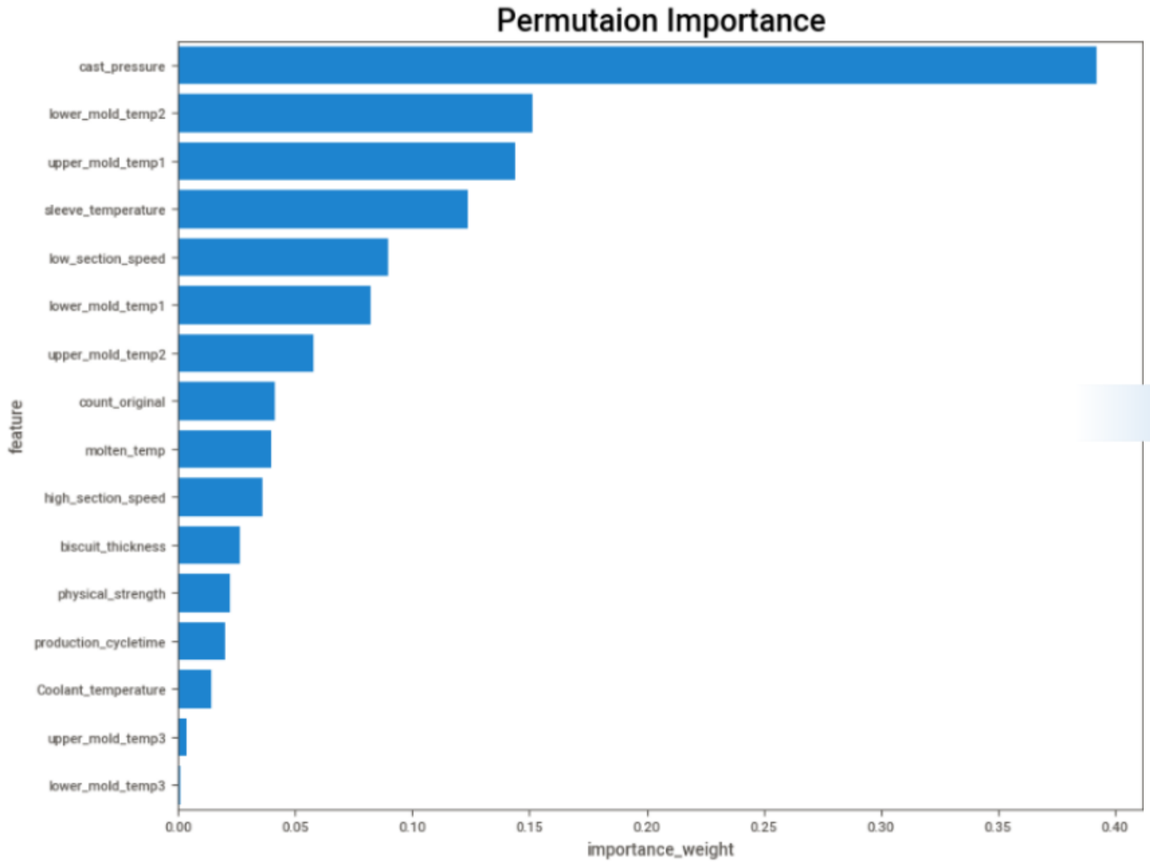




설비 작동 사이클시간
저속구간속도
상금형온도1
상금형온도2
하금형온도2
하금형온도2
냉각수 온도

데이터 모델링 및 시각화

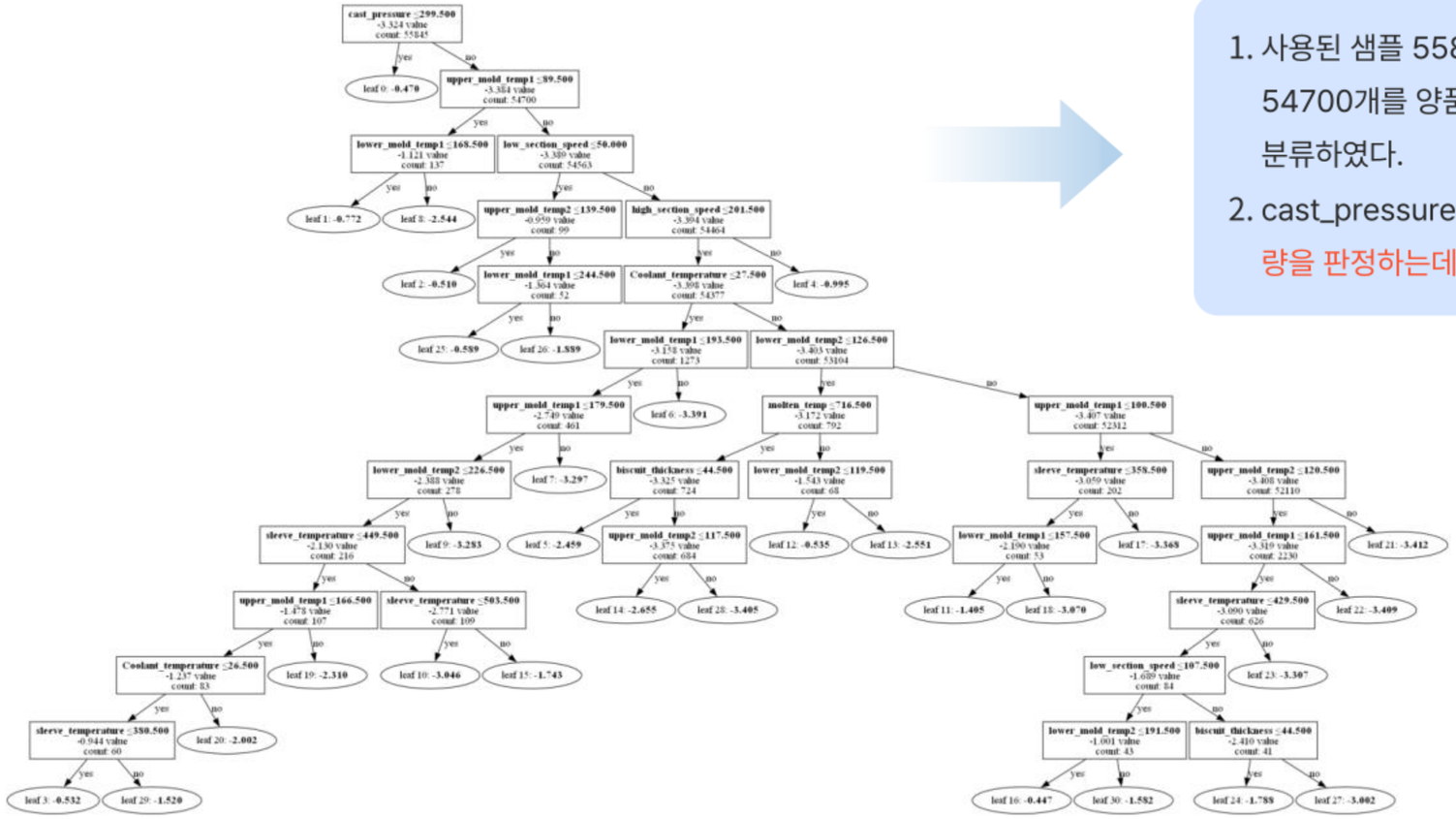
1. 모델의 Permutation Importance 확인



'cast_pressure'(주조압력)와 'lower_mold_temp2'(하금형온도2)
변수 중요도가 높다.
→ 해당 변수들이 주조 제품을 양품 or 불량으로 분류하는데 영향력이 높다.

데이터 모델링 및 시각화

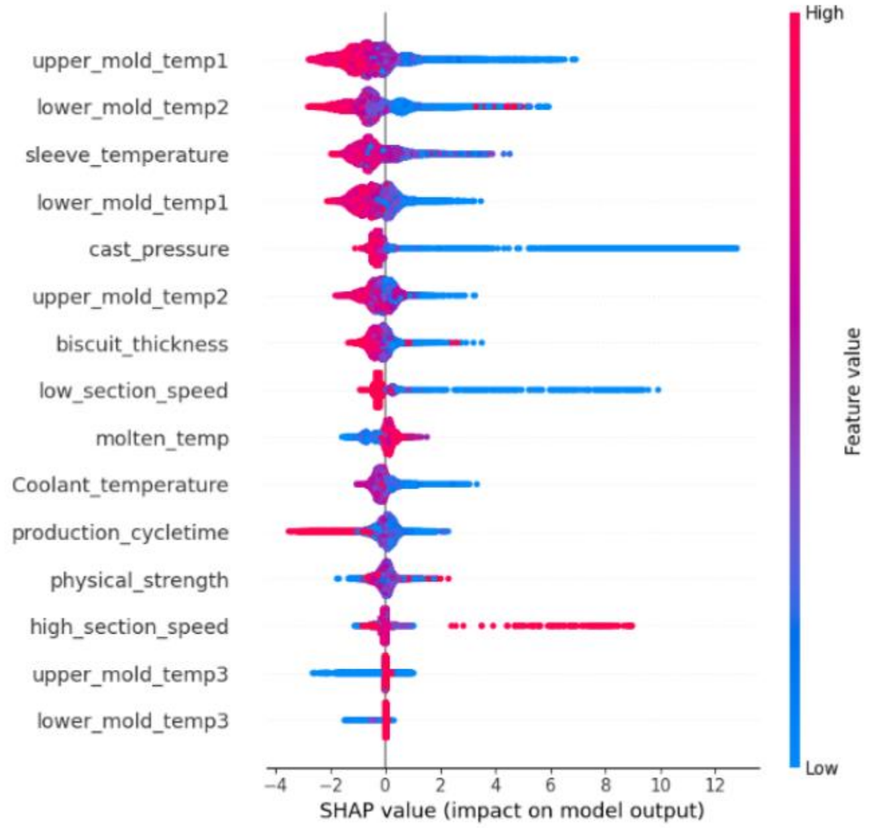
2. 모델 트리 시각화 및 최적화



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.

데이터 모델링 및 시각화

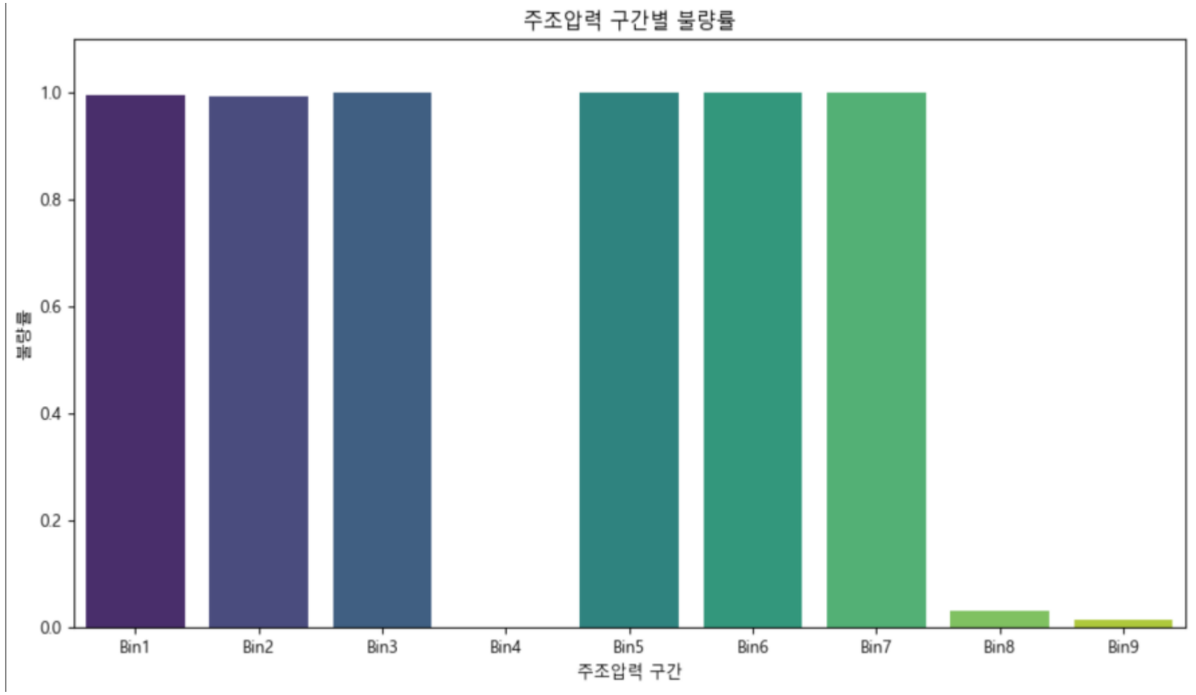
3. SHAP 시각화



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.

불량 발생 구간 확인

1. cast_pressure 구간 확인



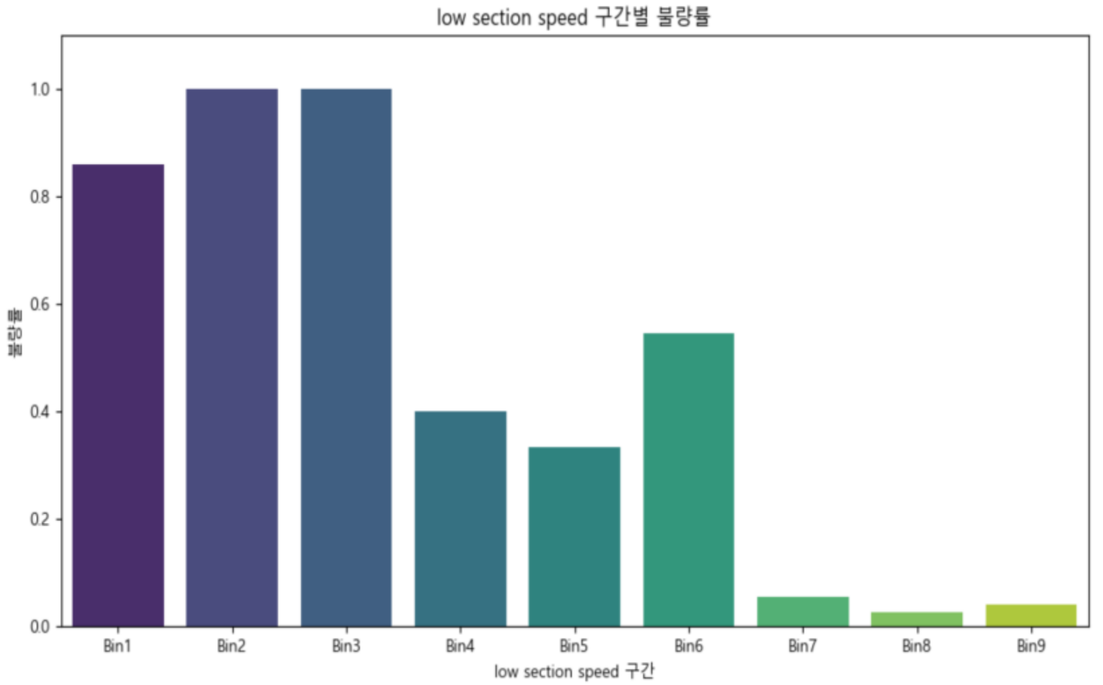
구간	구간 경계값	불량률
Bin1	143.00~164.67	99.51%
Bin2	164.67~186.33	99.28%
Bin3	186.33~208.00	100.00%
Bin4	208.00~229.67	nan%
Bin5	229.67~251.33	100.00%
Bin6	251.33~273.00	100.00%
Bin7	273.00~294.67	100.00%
Bin8	294.67~316.33	3.08%
Bln9	316.33~338.00	1.44%

▼

공장장님! 불량품의 개수를 줄이기 위해서는 주조압력을
Bin8(294.67MPa) 이상으로 유지해야 합니다!

불량 발생 구간 확인

2. 저속/고속구간속도 구간 확인



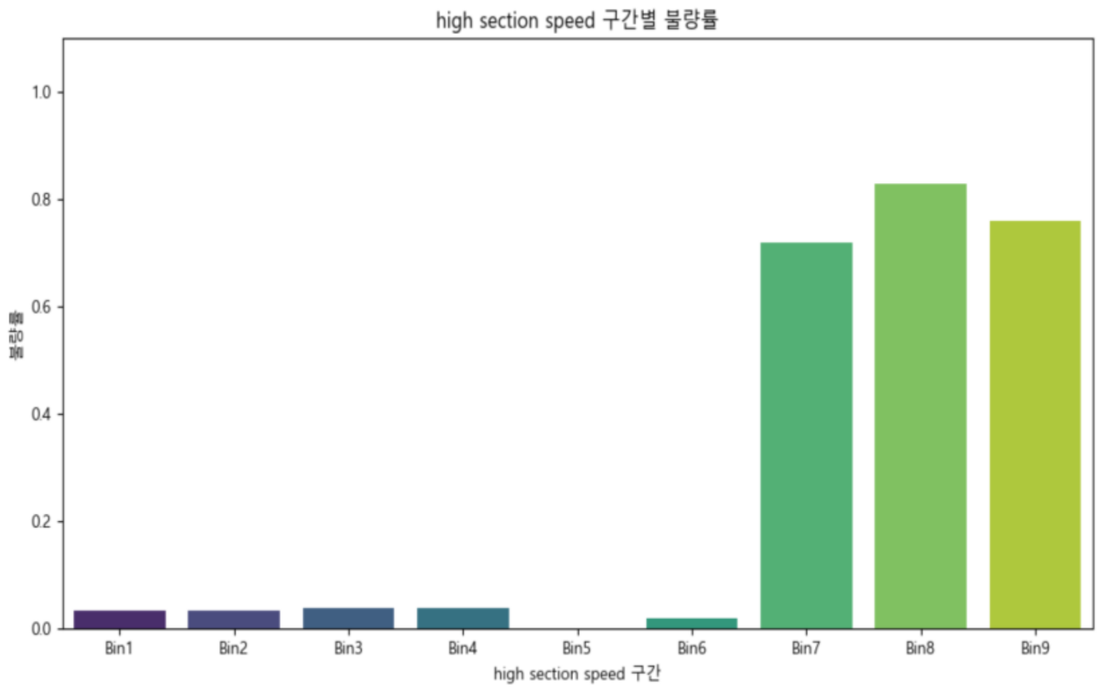
구간	구간 경계값	불량률
Bin1	0.00~15.56	85.96%
Bin2	15.56~31.11	100.00%
Bin3	31.11~46.67	100.00%
Bin4	46.67~62.22	40.00%
Bin5	62.22~77.78	33.33%
Bin6	77.78~93.33	54.35%
Bin7	93.33~108.89	5.39%
Bin8	108.89~124.44	2.65%
Bln9	124.44~140.00	3.95%

▼

공장장님! 불량품의 개수를 줄이기 위해서는 저속구간속도를
Bin7(93.33m/sec) 이상으로 유지해야 합니다!

불량 발생 구간 확인

2. 저속/고속구간속도 구간 확인



구간	구간 경계값	불량률
Bin1	101.00~117.33	3.34%
Bin2	117.33~133.67	3.37%
Bin3	133.67~150.00	3.67%
Bin4	150.00~166.33	3.77%
Bin5	166.33~182.67	0.00%
Bin6	182.67~199.00	1.92%
Bin7	199.00~215.33	71.74%
Bin8	215.33~231.67	82.86%
BIn9	231.67~248.00	75.86%

▼

공장장님! 불량품의 개수를 줄이기 위해서는 고속구간속도를 Bin6(199.00m/sec) 이하로 유지해야 합니다!

T-test

3. 양품집단과 불량집단간의 T-test 성능

	col	tvalue	pvalue
0	일자별 제품 생산 번호	-45.044308	0.000000e+00
1	용탕온도	3.786891	1.552251e-04
2	설비 작동 사이클시간	0.830063	4.065656e-01
3	제품생산 사이클 시간	-16.718089	3.974998e-60
4	저속구간속도	-17.048564	2.804018e-62
5	고속구간속도	7.086440	1.706243e-12
6	주조압력	-63.993068	0.000000e+00
7	비스켓 두께	3.641038	2.761401e-04
8	상금형온도1	-48.991533	0.000000e+00
9	상금형온도2	-51.139842	0.000000e+00
10	상금형온도3	1.702133	8.882696e-02
11	하금형온도1	-56.800771	0.000000e+00
12	하금형온도2	-56.687884	0.000000e+00
13	하금형온도3	-2.643416	8.248193e-03
14	슬리브온도	-18.661645	8.231163e-74
15	형체력	-8.474684	3.546059e-17
16	냉각수 온도	-39.398231	2.296088e-275
17	전자교반 가동시간	4.266051	2.046813e-05
18	양품불량판정	inf	0.000000e+00
19	금형코드	4.761839	2.002902e-06

양품 집단과 불량 집단 간 T-test

양품 집단과 불량 집단 간 T-test를 진행하고 그 결과를 확인한다.

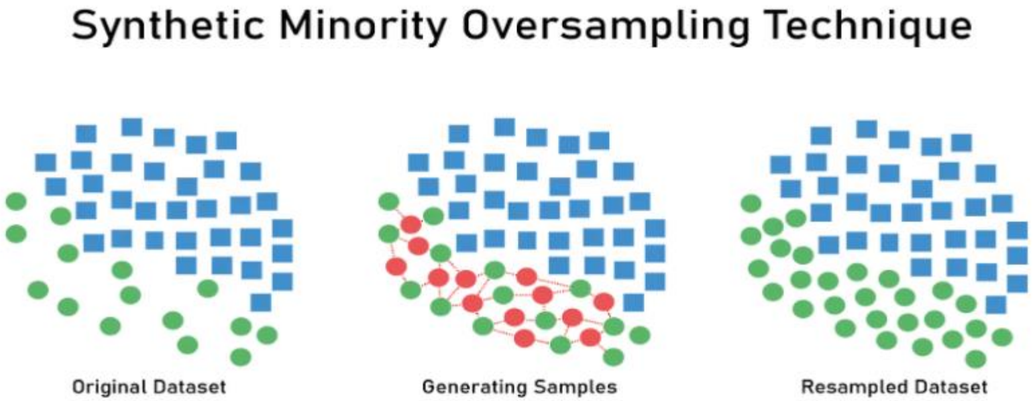
- T-test는 양품과 불량 데이터 간 차의가 유의미한 변수를 확인하는 것
- H_0(귀무가설) : '양품과 불량 집단의 평균 차이가 없다.'
- H_1(대립가설) : '양품과 불량 집단의 평균 차이가 있다.'

T-test 결론

T-test 결과, p-value를 반환하게 되며, 그 값이 0.1 미만일 경우 귀무가설 기각, 대립가설을 채택한다. 해당 데이터는 설비 작동 사이클시간 (facility_operation_cycleTime)를 제외하고, p-value 값이 0.1 미만이므로 대립가설을 채택한다.

학습모형 개발

4. SMOTE 기법을 통한 모델링



데이터 불균형 문제 해결

결측치 및 이상치 탐지를 통해 확보한 양품/불량품 데이터에는 양적 불균형이 존재하고 있다. 따라서, 데이터셋 불균형을 해소하기 위해 SMOTE 기법을 이용하여 오버샘플링을 수행하였고 데이터셋의 균형을 맞추었다.

학습모형 개발

4. SMOTE 기법을 통한 모델링

Model	Accuracy	Specificity	Recall	Precision	F1-score
Logistic Regression	Split 1 0.9140444215897628	Split 4 0.957753164556962	Split 1 0.8717563291139241	Split 4 0.9536659436008676	Split 1 0.9102474082028831
AdaBoost	Split 1 0.9529874804691363	Split 4 0.973615506329114	Split 1 0.9331091772151898	Split 4 0.972215279513455	Split 1 0.9520330945414185
SVM	Split 4 0.9163189019204525	Split 4 0.9163189019204525	Split 1 0.8748417721518987	Split 4 0.9554305495456512	Split 4 0.9125663863115043
Random Forest	Split 5 0.9964992780997212	Split 3 0.9943435781812429	Split 5 0.9987342272853131	Split 3 0.9943654202293235	Split 5 0.9965071534287124

주조 데이터를 분석하는 궁극적인 목적은 '불량품 판별'이다.
따라서 불량품을 정확하게 불량품으로 판별하는 평가지표인 특이도(Specificity)를 평가지표로 채택하였다.
특이도 측면에서는 Random Forest 모형이 가장 좋은 결과를 보여주고 있으므로 이 모형을 양불 판단 모형으로 채택하였다.

주조 데이터 스마트 생산 관리 시스템

주조 공정 data 수집

양품 Data

최적의 공정 환경 유지

양품 총량 증가

총 생산량 Data

불량품

AUTO

공정환경 조절

불량유형 분석(라벨링)

양품총량	총 생산량	441
	현재 양품	22
422	총 불량품	19
	불량률	4.31%
	불량 원인 추정	주조압력 확인 요망!

사전기획 관점에서 프로젝트 결과

1. 불량율이 왜 발생했는지 변수 분석을 통해 주조압력, 저속구간속도, 고속구간속도가 불량률에 영향을 미치는 변수라는 것을 파악한 뒤, 구간 분석을 통해 공정 요소 관리에 대한 인사이트를 제공하였다.
2. 'SMOTE 모델을 통한 스마트 생산관리 시스템' 이라는 비즈니스 모델을 통해 주조 공정에서 발생하는 공정 문제를 원격으로 제어함으로써 불량품을 관리할 수 있는 플랫폼을 제공하였다.
3. 추후 계획: 불량율 감소를 위한 추가적인 조치나 지속적인 모니터링 계획을 수립할 예정이다.

자체평가

10점/10점

아쉬운점

1. 원인 파악의 한계: 불량율이 감소되었지만, 모든 원인을 파악하고 해결하지 못했습니다. 추가적인 원인 분석과 개선이 필요하다.
2. 주조공정에 대한 이해와 배경지식이 부족한 상태에서 데이터 분석을 진행하였기 때문에 놓친 부분이 있을수 있을 것이다.

잘한 점

1. 불량율 감소: 프로젝트 결과로 불량률을 줄일 수 있는 공정 최적화를 위한 주요 변수 구간을 제시하였다.
2. 팀 협력: 팀원들 간의 협력이 원활하였고, 함께 문제를 해결하고 개선 방안을 모색하는 데 최선을 다하였다.
3. 품질 의식 강화: 프로젝트를 통해 조직 내부의 품질 의식이 높아졌다. 이는 향후 품질 관리에 더 많은 주의를 기울일 수 있을 것이다.

감사합니다