

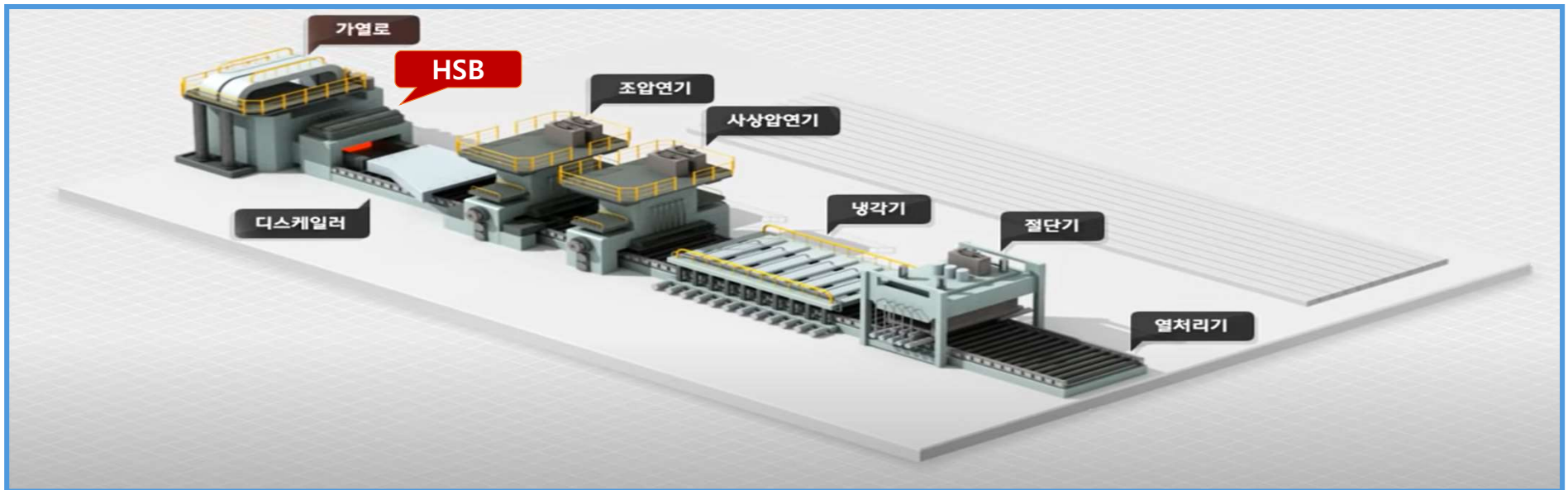


압연 공정 Scale 불량 원인 분석 및 개선안 도출

C반 김민정

■ 분석 배경

- 고객사로부터 Scale 불량 증가 이슈 확인함
- 1차 원인 분석 결과 압연공정에서 Scale 불량률이 급증하여, 데이터 분석을 통해 근본 원인을 찾아 개선 기회 도출하고자 함



■ 발생 현황

발생원인	압입함	Scratch	두께분석	Scale	계
발생률(%)	1.3	0.5	0.4	5.0	7.2%

- 잠재적 인자 선정
 - 현업 엔지니어들과 협의를 통해 하기 잠재적 원인을 도출함

Scale 발생	가열로 가열대 온도	가열로 균열대 온도	가열로 추출 온도	Hot Scale Breaker	사상 압연 온도	압연간 Descaling 횟수	판두께
없음 ↕ 발생	저 ↕ 고	저 ↕ 고	저 ↕ 고	적용 ↕ 미적용	저 ↕ 고	증가 ↕ 감소	후판 ↕ 박판

- 분석 방향
 - 다양한 분류 모델을 활용하여 Scale 불량 발생에 영향을 주는 인자를 도출
 - 분석을 통해 도출된 영향인자와 사전 검토한 잠재인자 간의 일치여부 및 방향성 등을 확인하고 정리
 - 최종적으로 모델 평가를 통해 모델 선정을 진행하고 개선안을 도출함

목적	분석방법	주요내용	담당자	일정	비고
데이터 특성 파악	기초통계 분석	데이터의 타입, 평균, 표준편차, 최소값, 최대값 등의 기본 정보 확인	김민정	8/16	
	결측치 분석	데이터에 결측치가 있는지 확인 후 처리	김민정	8/16	
전체 데이터의 분포 특성 및 변수 간의 관련성 확인	Histogram 분석	목표변수와 연속형 변수간 구간별 분포 현황과 패턴 확인	김민정	8/16	
	Box Plot 분석	Box plot을 이용하여 데이터에 이상치가 있는지 확인 후 처리	김민정	8/16	
	Crosstab 분석	목표변수와 범주형 변수간 빈도 및 비율 확인	김민정	8/16	
	카이제곱 검정	목표변수와 범주형 변수간 통계적 유의성 검정	김민정	8/16	
잠재인자 (Vital Few) 탐색	Decision Tree	Decision Tree를 통해 변수 별 중요도 확인 후 잠재인자 도출	김민정	8/16	
	Random Forest	Random Forest를 통해 변수 별 중요도 확인 후 잠재인자 도출	김민정	8/16	
	Gradient Boosting	Gradient Boosting을 통해 변수 별 중요도 확인 후 잠재인자 도출	김민정	8/16	
예측 모델링 진행	Logistic Regression	도출된 잠재인자를 통하여 로지스틱 회귀모델 생성	김민정	8/17	
	Decision Tree	도출된 잠재인자를 통하여 분류 DT모델 생성	김민정	8/17	
	Random Forest	도출된 잠재인자를 통하여 분류 RF모델 생성	김민정	8/17	
	Gradient Boosting	도출된 잠재인자를 통하여 분류 GB모델 생성	김민정	8/17	
	SVM	도출된 잠재인자를 통하여 SVM 모델 생성	김민정	8/17	
	KNN	도출된 잠재인자를 통하여 KNN 모델 생성	김민정	8/17	
모델 평가	모델평가함수	AUC, Precision, Recall, F1 Score 확인 및 최종 모델평가	김민정	8/17	

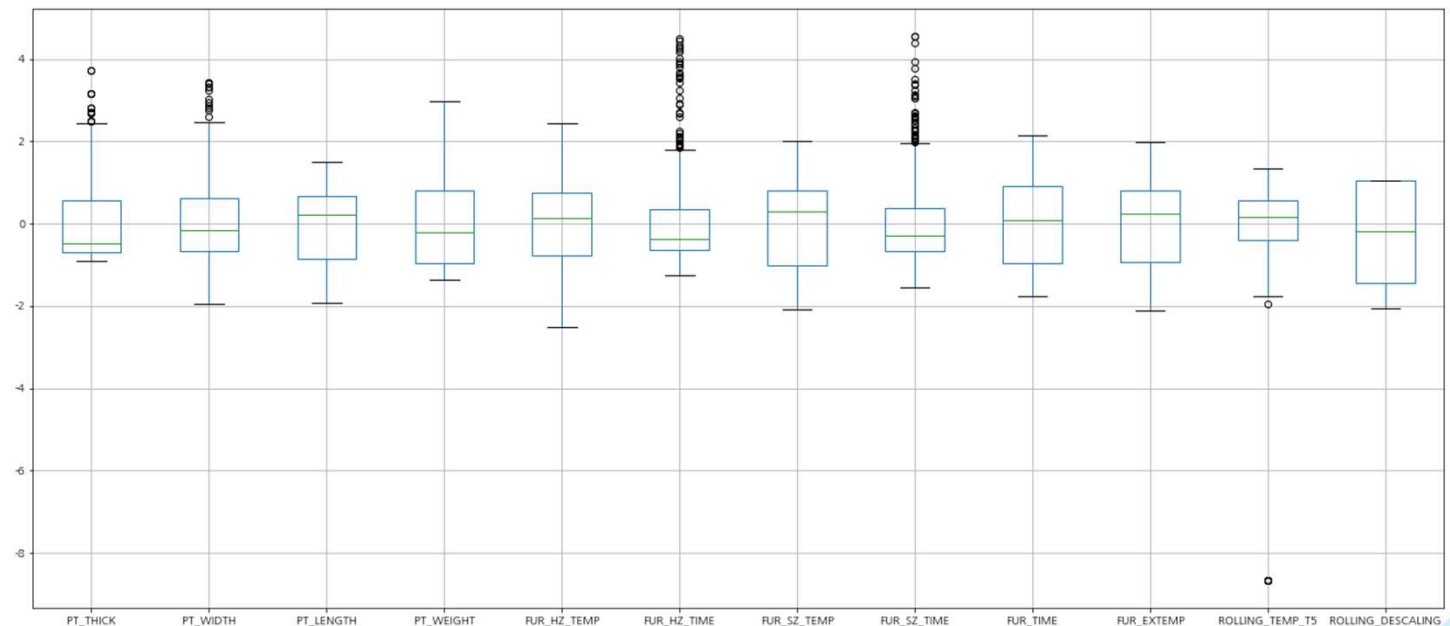
- 결측치 처리
 - 결측치가 없어서 따로 처리하지 않음
- 이상치 처리
 - 이상치 전체 확인을 위해 Standard Scaling 후 Box plot 생성
 - PT_THICK(Plate 두께), PT_WIDTH(폭), FUR_HZ_TIME(가열로 가열대 시간), FUR_SZ_TIME(가열로 균열대 시간), ROLLING_TEMP(압연온도)에서 이상치가 확인되나, 히스토그램을 통해 목표변수와 관계 확인 후 처리예정

<결측치 확인>

```

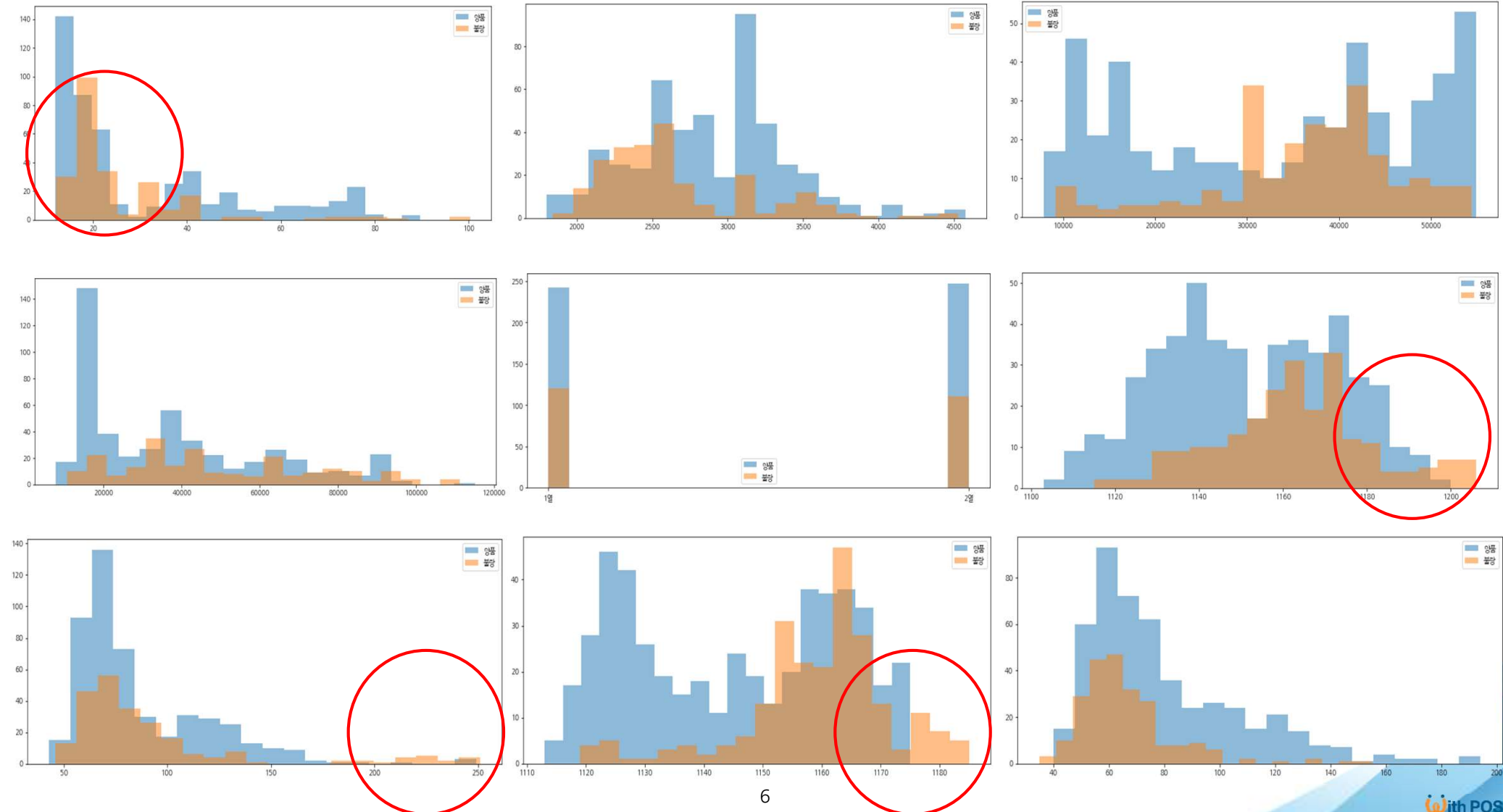
PLATE_NO      0
ROLLING_DATE  0
SCALE         0
SPEC          0
STEEL_KIND    0
PT_THICK      0
PT_WIDTH      0
PT_LENGTH     0
PT_WEIGHT     0
FUR_NO        0
FUR_NO_ROW    0
FUR_HZ_TEMP   0
FUR_HZ_TIME   0
FUR_SZ_TEMP   0
FUR_SZ_TIME   0
FUR_TIME      0
FUR_EXTEMP    0
ROLLING_TEMP_T5 0
HSB           0
ROLLING_DESCALING 0
WORK_GR       0
dtype: int64
    
```

<Box plot 확인>



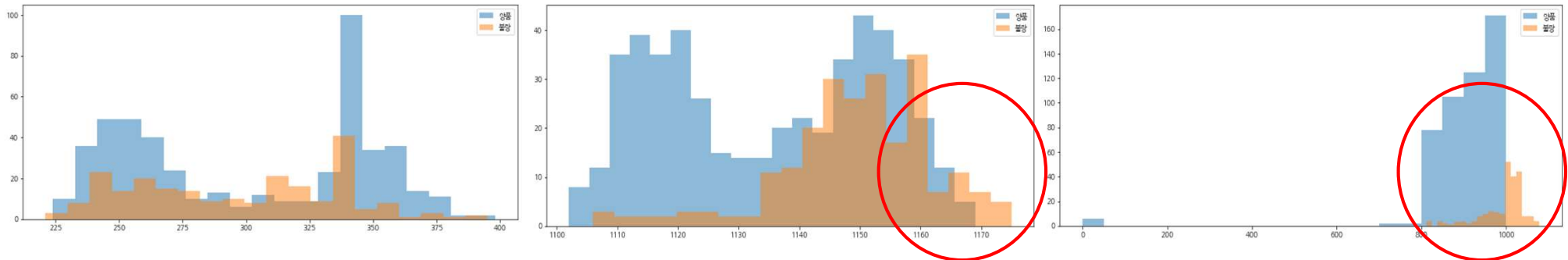
■ Histogram을 통한 목표변수 - 연속형 변수 간 관계 파악

- FUR_HZ_TEMP, FUR_SZ_TEMP, FUR_EXTEMP, ROLLING_TEMP_T5: 온도변수는 특정온도에서 불량률이 높아 Vital Few로 고려함



■ Histogram을 통한 목표변수 - 연속형 변수 간 관계 파악

- PT_THICK, PT_WIDTH, PT_LENGTH, PT_WEIGHT: 고객 needs에 따라 변경되는 수치로 불량률과 직접적인 연관성을 찾기 어렵다고 판단하여 Vital Few에서 제외



■ Cross tab 및 카이제곱 검정을 통한 목표변수 - 범주형 변수 간 관계 파악

- 가열로 호기에 따른 불량률: 카이제곱 검정 결과 p-value가 0.05보다 커서 유의수준 5%에서 불량률 차이 없다고 판단
- 강종에 따른 불량률: 카이제곱 검정 결과 p-value가 0.05보다 작아 유의수준 5%에서 불량률 차이가 있다고 판단 > **Vital Few** 후보

FUR_NO	1호기	2호기	3호기
SCALE			
불량	73	70	88
양품	167	167	155

FUR_NO	1호기	2호기	3호기
SCALE			
불량	0.304	0.295	0.362
양품	0.696	0.705	0.638

```
chi-square test
chisq: 2.915
p: 0.233
degree pf freedom: 2
expected value:
[[ 77.    76.038  77.962]
 [163.   160.962 165.038]]
```

STEEL_KIND	C0	C1	C3	T0	T1	T3	T5	T7	T8
SCALE									
불량	212	1	1	2	2	0	2	6	5
양품	291	0	6	14	16	2	41	29	90

STEEL_KIND	C0	C1	C3	T0	T1	T3	T5	T7	T8
SCALE									
불량	0.421	1.0	0.143	0.125	0.111	0.0	0.047	0.171	0.053
양품	0.579	0.0	0.857	0.875	0.889	1.0	0.953	0.829	0.947

```
chi-square test
chisq: 73.654
p: 0.000
degree pf freedom: 1
expected value:
[[163.066  65.934]
 [343.934 139.066]]
```

- **Cross tab 및 카이제곱 검정을 통한 목표변수 - 범주형 변수 간 관계 파악**
 - HSB 적용/미적용에 따른 불량률: 카이제곱 검정을 진행하지 않아도 미적용시 100% 불량인 점을 확인함.
 - Descaling 횟수에 따른 불량률: 횟수가 홀수인 데이터는 100%불량이므로, 카이제곱 검정시에는 제외하고 짝수인 데이터에 대하여 검정 진행함.
 - ▶ p-value가 0.05보다 낮으므로 유의수준 5%에서 불량률에 차이가 있음을 확인함. (**Vital Few 후보**)
 - 작업조에 따른 불량률: 카이제곱 검정 진행 시 p-value가 0.05보다 크므로 유의수준 5%에서 불량률 차이 없는 것으로 판단함.

HSB	미적용	적용
SCALE		
불량	33	198
양품	0	489

HSB	미적용	적용
SCALE		
불량	1.0	0.288
양품	0.0	0.712

ROLLING_DESCALING	5	6	7	8	9	10
SCALE						
불량	5	26	15	99	20	66
양품	0	150	0	109	0	230

ROLLING_DESCALING	5	6	7	8	9	10
SCALE						
불량	1.0	0.148	1.0	0.476	1.0	0.223
양품	0.0	0.852	0.0	0.524	0.0	0.777

```

chi-square test
chisq: 60.196
p: 0.000
degree of freedom: 2
expected value:
[[ 47.812  58.5    82.688]
 [122.188 149.5   211.312]]
  
```

WORK_GRP	1조	2조	3조	4조
SCALE				
불량	67	45	54	65
양품	122	120	118	129

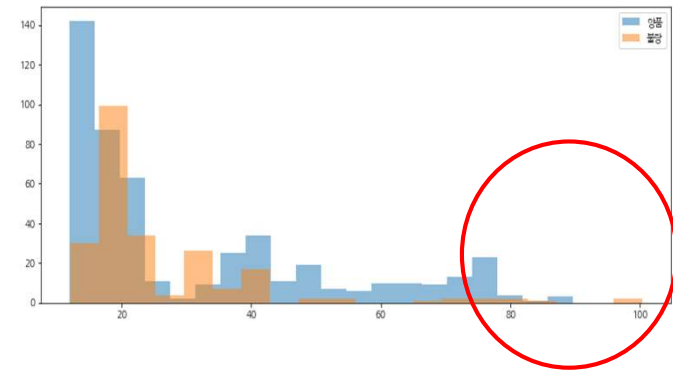
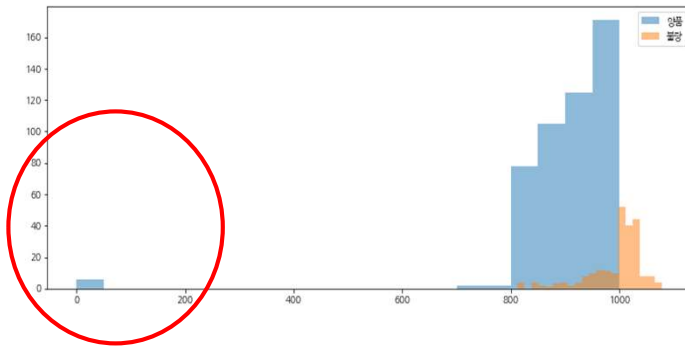
WORK_GRP	1조	2조	3조	4조
SCALE				
불량	0.354	0.273	0.314	0.335
양품	0.646	0.727	0.686	0.665

```

chi-square test
chisq: 2.868
p: 0.412
degree of freedom: 3
expected value:
[[ 60.466  52.747  54.034  61.753]
 [127.534 111.253 113.966 130.247]]
  
```


■ 이상치 제거

- 압연온도의 경우 양품임에도 불구하고 0도 부근의 데이터가 있어 이상치로 판단하여 0-40도 데이터 제외함.
- 주문두께의 경우 90 이상이 모두 불량이지만 고객 needs를 한정하기에는 무리가 있다고 판단하여 해당 데이터는 제외함.



■ 추가 Column 처리

- 앞선 분석을 통해 분석에 필요한 데이터가 아니라고 판단된 설명변수는 제외함.

```
df_raw.drop(['PLATE_NO', 'ROLLING_DATE', 'FUR_NO', 'FUR_NO_ROW', 'WORK_GR'], axis=1, inplace=True)
```

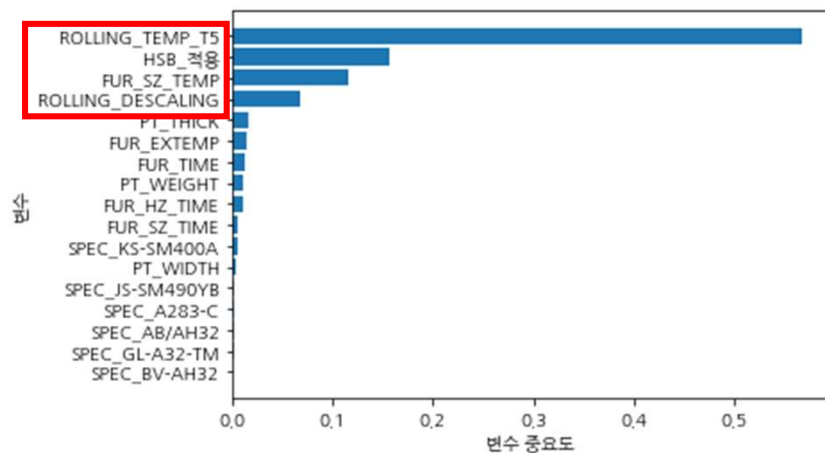
	SCALE	SPEC	STEEL_KIND	PT_THICK	PT_WIDTH	PT_LENGTH	PT_WEIGHT	FUR_HZ_TEMP	FUR_HZ_TIME	FUR_SZ_TEMP	FUR_SZ_TIME	FUR_TIME	FUR_EXTEMP	ROLLING_TEMP_T5	HSB	ROLLING_DESCALING
0	양품	AB/EH32-TM	T1	32.25	3707	15109	14180	1144	116	1133	59	282	1125	934	적용	8
1	양품	AB/EH32-TM	T1	32.25	3707	15109	14180	1144	122	1135	53	283	1120	937	적용	8
2	양품	NV-E36-TM	T8	33.27	3619	19181	18130	1129	116	1121	55	282	1106	889	적용	8
3	양품	NV-E36-TM	T8	33.27	3619	19181	18130	1152	125	1127	68	316	1113	885	적용	8
4	양품	BV-EH36-TM	T8	38.33	3098	13334	12430	1140	134	1128	48	314	1118	873	적용	8
...
715	불량	NK-KA	C0	20.14	3580	38639	21870	1172	72	1164	62	245	1155	1005	적용	8
716	양품	NV-A32	C0	15.08	3212	48233	18340	1150	61	1169	61	238	1160	947	적용	10
717	양품	NV-A32	C0	16.60	3441	43688	19590	1169	65	1163	77	247	1152	948	적용	10
718	양품	LR-A	C0	15.59	3363	48740	80240	1179	86	1163	45	243	1154	940	적용	10
719	양품	GL-A32	C0	16.09	3400	54209	69840	1186	82	1169	45	239	1155	957	적용	10

712 rows × 16 columns

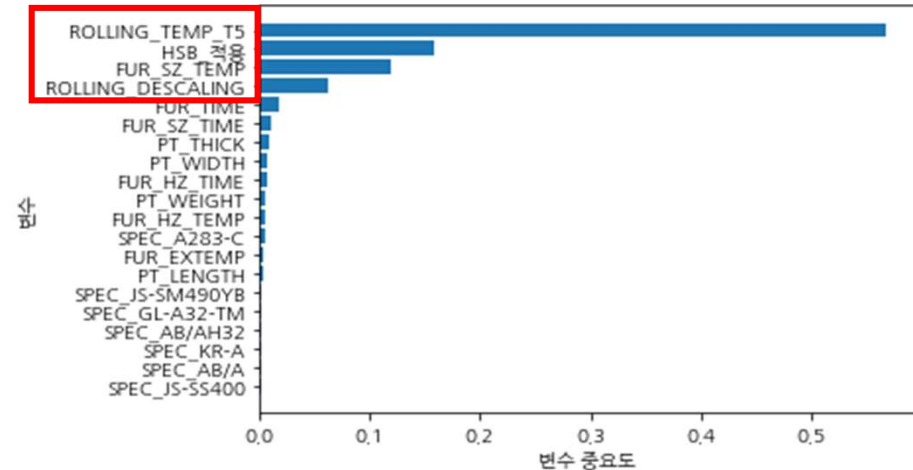
DT/RF/GB를 통한 변수중요도 확인

- 세가지 모델을 통해 변수중요도를 확인한 결과 4가지의 공통 Vital Few 확인함

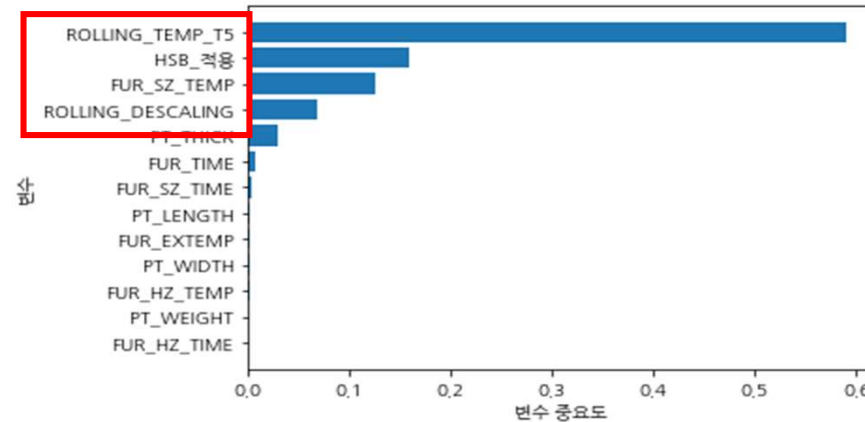
<Decision Tree>



<Random Forest>



<Gradient Boosting>



■ 최종 영향인자 도출

- 데이터 분석결과 압연온도, HSB 적용여부, 균열대 온도, Descaling 횟수가 주요 영향인자임을 알 수 있었음

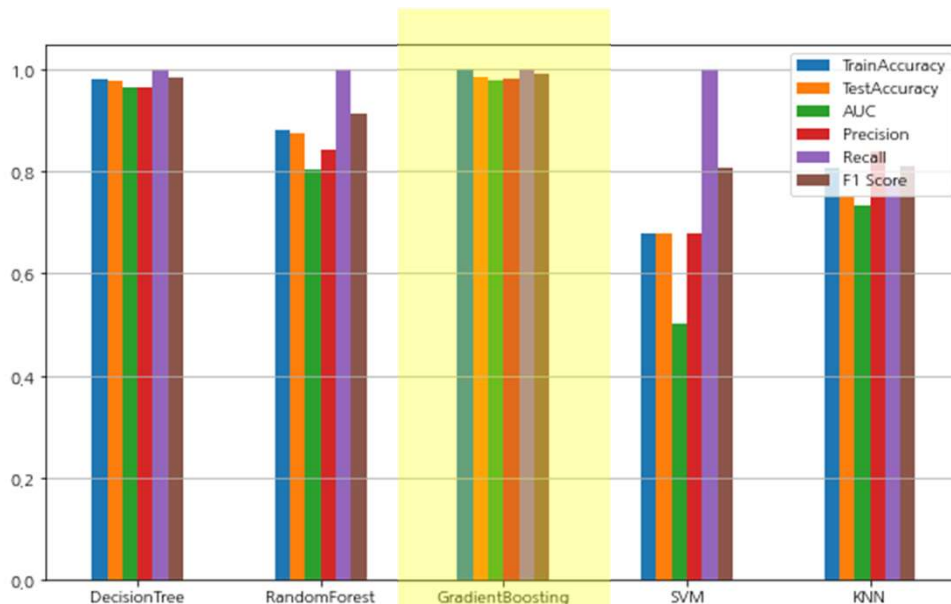
■ 모델간 정확도 비교

- Gradient Boosting > Decision Tree > Random Forest > KNN > SVM > Logistic Regression 순으로 예측률이 높음

모델 종류	Logistic Regression	Decision Tree	Random Forest	Gradient Boosting	SVM	KNN
정확도(%)	40.8	97.7	87.4	98.6	67.8	75.2

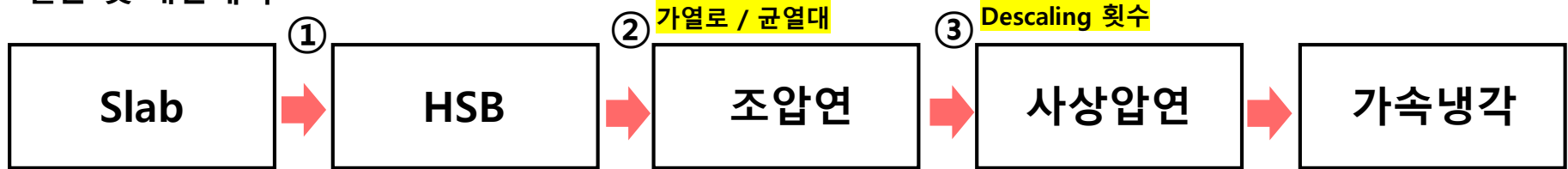
■ 모델 평가

- AUC / Precision / Recall / F1 Score 기준으로 GB > DT > RF > KNN > SVM 순으로 정확함을 알 수 있다.
- 따라서, 전체적인 모델 평가 지표의 밸런스가 가장 좋은 **Gradient Boosting**을 최종 모델로 선정함.



	TrainAccuracy	TestAccuracy	AUC	Precision	Recall	F1 Score
DecisionTree	0.980	0.977	0.964	0.967	1.000	0.983
RandomForest	0.882	0.874	0.804	0.843	1.000	0.915
GradientBoosting	1.000	0.986	0.978	0.980	1.000	0.990
SVM	0.679	0.678	0.500	0.678	1.000	0.808
KNN	0.807	0.752	0.734	0.838	0.786	0.811

▪ 결론 및 대안제시



① 후판공정 진행 시 Slab에서 HSB 공정은 필수적으로 진행이 필요함

② 가열로부터 압연 공정 진행까지 일정 온도 유지가 필요함

- 가열로 균열대 온도가 1150℃ 이상인 경우 압연온도를 1000℃ 이내로 맞추려고하여도 1000℃ 이상이 되어 scale 불량률이 현저히 높아질 것으로 예상함

③ 압연간 Descaling 횟수는 선단부, 후단부가 모두 진행 되어야 불량 발생 가능성이 감소할 것으로 예상됨.

- 데이터 분석 결과 홀수 번째에서는 항상 불량이 발생하는 것을 알 수 있음.
- 또한, Descaling을 6회 진행하였을때 불량률이 가장 낮은 것이 확인되나, 기존 잠재인자에 따르면 Descaling 횟수와 불량 발생이 반비례 관계임으로 해당 부분은 현업 내 기기 검사가 필요할 것으로 생각됨.

- ✓ 2번째 종합실습은 1번째 종합실습 대비 모델링 진행이 익숙해져 조금 더 수월하게 진행할 수 있었습니다.
- ✓ 이번 분석에서도 정확한 데이터 분석 진행을 위해서는 도메인 지식이 중요함을 한번 더 느끼게 되었습니다. 공정에 대한 지식이 부족하여 분석 결과에 대한 해석이 어려웠습니다.
- ✓ 분류 모델 평가시에는 AUC / Precision / Recall / F1 Score에 대한 이해가 뒷받침 되어야 모델 성능에 대한 정확한 평가가 진행될 수 있음을 알게 되었습니다.

- 1페이지 배경: <https://m.post.naver.com/viewer/postView.naver?volumeNo=33977027&memberNo=41266020&vType=VERTICAL>
- 2페이지 그림: <https://changwan1202.tistory.com/95>