종합실습 2

불량 발생 원인 파악 및 결과 해석

Scale 불량 발생 증가 문제 개선 기회 도출

A반 3조 김준오

INDEX

0 과제 정의

분석 배경

OO공장에의 고객사에서 최근 들어 :Scale 불량 발생 증가"라는 이슈가 발생했다.
그 원인을 분석해 본 결과 압연 공정에서 Scale 불량이 급증한 것을 확인할 수 있었다.

그래서, 데이터를 수집하여 다양한 분석을 통해 불량 발생의 근본 원인을 찾고
결과를 해석하여 개선 기회를 도출한다.

1 잠재 인자 설정

1. 가열로 가열대 온도(FUR_HZ_TEMP) : `저 - 고`
2. 가열로 균열대 온도(FUR_SZ_TEMP) : `저 - 고`
3. 가열로 추출 온도(FUR_EXTEMP) : `저 - 고`
4. Hot Scale Breaker(HSB) : `적용 - 미적용`
5. 사상 압연 온도(ROLLING_TEMP_T5) : `저 - 고`
6. 압연간 Descaling 횟수(ROLLING_DESCALING) : `증가 - 감소`
7. 판두께(PT_THICK) : `후 - 박`

Scale 발생 : `없음 - 발생`

2 데이터 처리 – 결측치, 데이터 타입 확인

	PLATE_NO	ROLLING_DATE	SCALE	SPEC	STEEL_KIND	PT_THK	PT_WDTH	PT_LTH	PT_WGT	FUR_NO	...	FUR_HZ_TEMP	FUR_HZ_TIME	FUR_SZ_T
0	PB562774	2008-08-01:00:00:15	양품	AB/EH32-TM	T1	32.25	3707	15109	14180	1호기	...	1144	116	
1	PB562775	2008-08-01:00:00:16	양품	AB/EH32-TM	T1	32.25	3707	15109	14180	1호기	...	1144	122	
2	PB562776	2008-08-01:00:00:59	양품	NV-E36-TM	T8	33.27	3619	19181	18130	2호기	...	1129	116	
3	PB562777	2008-08-01:00:01:24	양품	NV-E36-TM	T8	33.27	3619	19181	18130	2호기	...	1152	125	
4	PB562778	2008-08-01:00:01:44	양품	BV-EH36-TM	T8	38.33	3098	13334	12430	3호기	...	1140	134	

5 rows × 21 columns

```
SCALE              0
SPEC               0
STEEL_KIND         0
PT_THK             0
PT_WDTH            0
PT_LTH             0
PT_WGT             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
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 720 entries, 0 to 719
Data columns (total 19 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   SCALE              720 non-null    object
 1   SPEC               720 non-null    object
 2   STEEL_KIND         720 non-null    object
 3   PT_THK             720 non-null    float64
 4   PT_WDTH            720 non-null    int64
 5   PT_LTH             720 non-null    int64
 6   PT_WGT             720 non-null    int64
 7   FUR_NO             720 non-null    object
 8   FUR_NO_ROW         720 non-null    int64
 9   FUR_HZ_TEMP        720 non-null    int64
 10  FUR_HZ_TIME        720 non-null    int64
 11  FUR_SZ_TEMP        720 non-null    int64
 12  FUR_SZ_TIME        720 non-null    int64
 13  FUR_TIME           720 non-null    int64
 14  FUR_EXTEMP         720 non-null    int64
 15  ROLLING_TEMP_T5    720 non-null    int64
 16  HSB                720 non-null    object
 17  ROLLING_DESCALING  720 non-null    int64
 18  WORK_GR            720 non-null    object
dtypes: float64(1), int64(12), object(6)
memory usage: 107.0+ KB
```

: Dtype이 object인데, int 형으로 되어있는 자료형이 있다.
이는 이후 데이터 전처리 과정에서 처리할 예정이다.

2 데이터 처리 – 결측치, 데이터 타입 확인

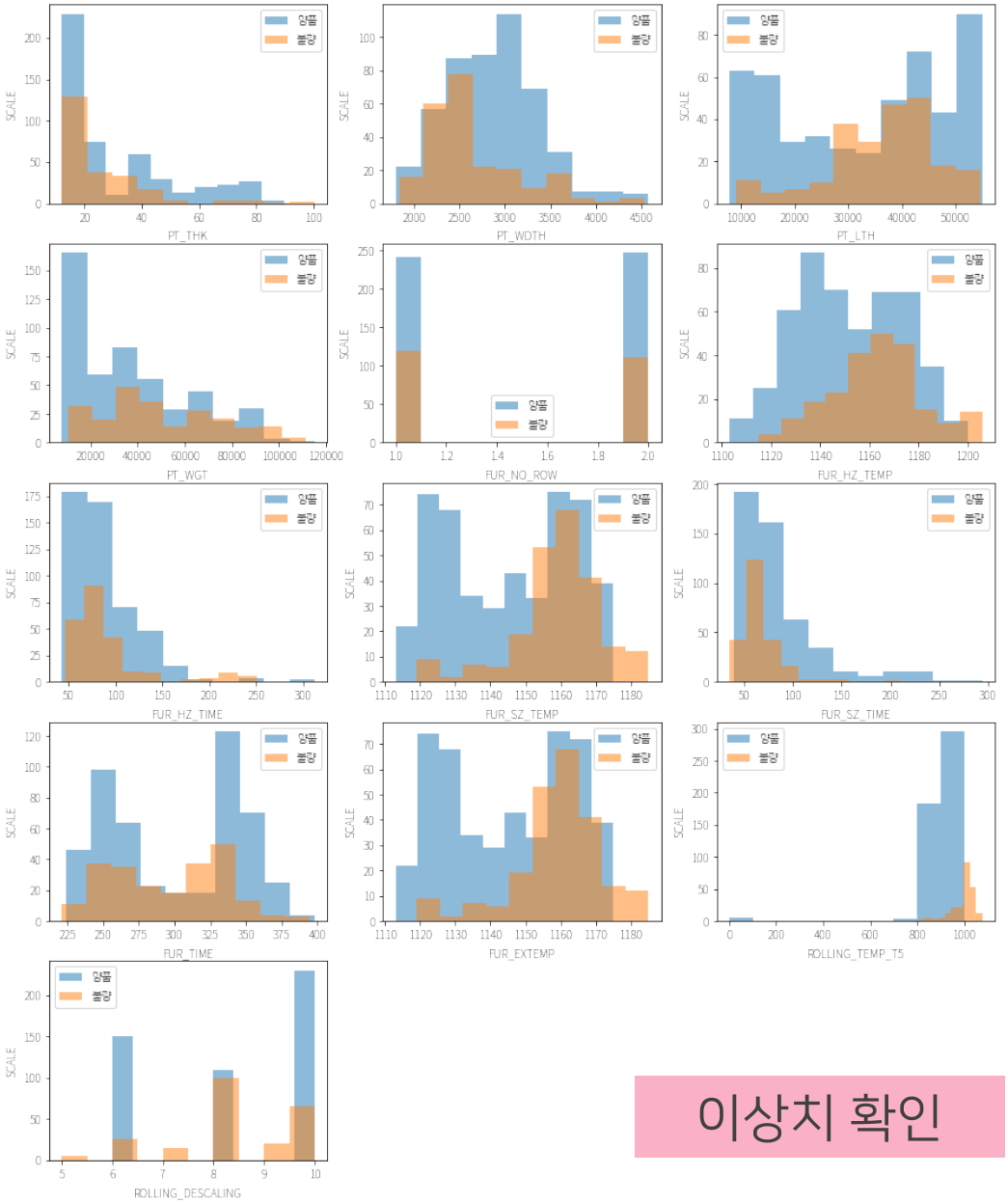
```
In [6]:  1  len(list(df_raw["SPEC"].unique()))
Out[6]: 66
```

: 66 종류의 제품 규격이 존재한다.
(너무 많은 카테고리는 분류에 방해가 될 수 있으므로 삭제하는 것을 고려해야한다.)

2 데이터 처리 – 이상치 확인

ROLLING_TEMP_T5 < 200 인
부분에 6개의 이상치가 존재했다.
→ 제거

이상치라 판단한 이유?
: 자료를 조사해 본 결과, "사상압연의 경우는 일반적으로 850~870℃정도이며, 탄소량이 낮은 냉연재의 경우는 900℃ 이상에서 사상압연을 완료해야 한다."라는 내용을 확인할 수 있었다. 따라서 사상압연의 온도가 200도 보다 낮으므로 이상치로 판단할 수 있었다.



이상치 확인

2 데이터 처리 – 이상치 확인

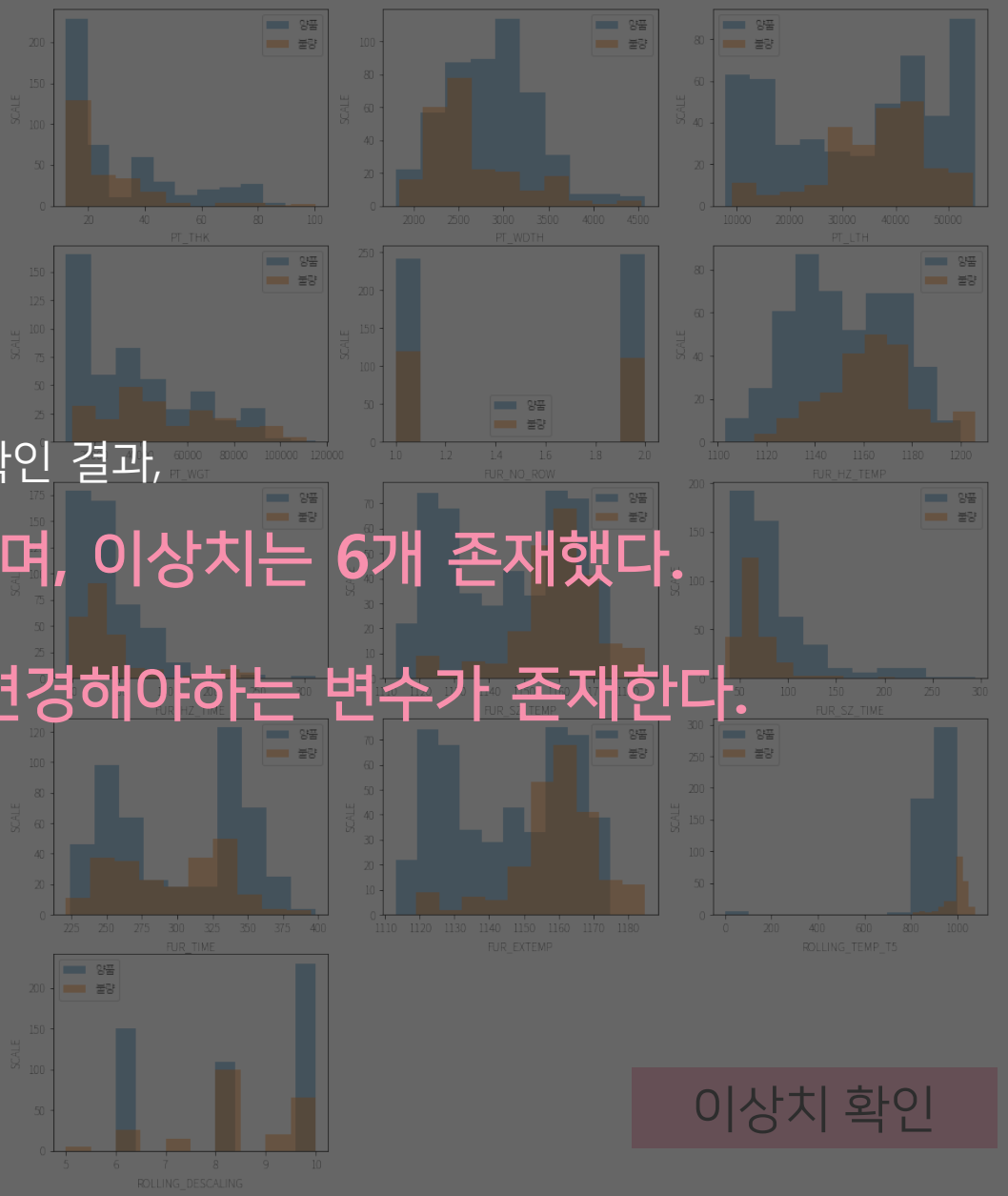
ROLLING_TEMP_T5 < 200 인
부분에 6개의 이상치가 존재했다.

→ 제거

데이터 확인 결과,

1. 결측치가 존재하지 않으며, 이상치는 6개 존재했다.

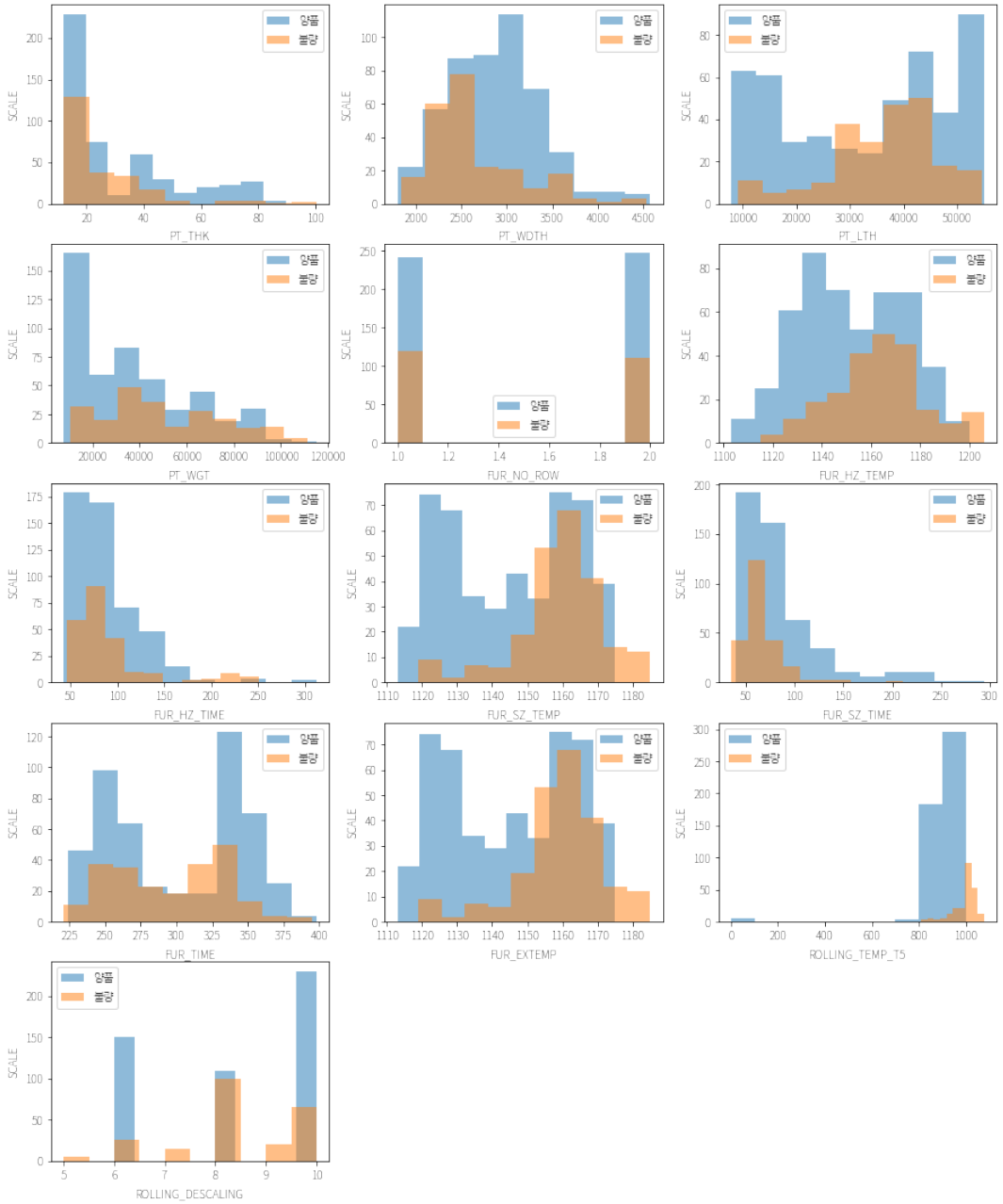
2. 데이터 타입을 **object** 로 변경해야하는 변수가 존재한다.



이상치 확인

3 탐색적 분석 : 연속형 변수

히스토그램 (연속형 변수)



3 탐색적 분석 – 연속형 변수 : 가설 검증

1. 가열로 가열대 온도(FUR_HZ_TEMP) : `저 - 고`. **(가설 O)**
2. 가열로 균열대 온도(FUR_SZ_TEMP) : `저 - 고` **(가설 O)**
3. 가열로 추출 온도(FUR_EXTEMP) : `저 - 고` **(가설 O)**
4. Hot Scale Breaker(HSB) : `적용 - 미적용` - 다음과정에서 확인

Scale 발생 : `없음 - 발생`

5. 사상 압연 온도(ROLLING_TEMP_T5) : `저 - 고` **(가설 O)**
6. 압연간 Descaling 횟수(ROLLING_DESCALING) : `증가 - 감소` (가설과 차이 존재)
7. 판두께(PT_THICK) : `후 - 박` **(가설 O)**

3 탐색적 분석 – 연속형 변수 : 가설 검증

8 기타 설명변수

: PT_WIDTH 가 작을수록 불량률이 증가하는 추세를 확인할 수 있었다.

: PT_LTH 가 중간 크기일 때, 가장 큰 불량률을 보이는 추세를 확인할 수 있었다.

: PT_WGT 가 클수록, 불량률이 증가하는 추세를 확인할 수 있었다.

: FUR_NO_ROW 는 1,2,3호기를 뜻하는 범주형 변수로, 불량률은 매우 유사하다.

: FUR_HZ_TIME 이 170을 넘어가는 순간, 불량률이 크게 증가하는 추세를 확인할 수 있었다.

: FUR_SZ_TIME 의 불량률은 70에서 최대값을 보인다.

: FUR_TIME 가 중간 크기일 떄, 가장 큰 불량률을 보이는 추세를 확인할 수 있었다.

3 탐색적 분석 – 범주형 변수

1. STEEL_KIND

```
1 fun_print_crosstab(df_raw, "STEEL_KIND")
```

```
STEEL_KIND    C0  C1  C3  T0  T1  T3  T5  T7  T8
SCALE
0            289   0   6  13  16   2  39  29  89
1            212   1   1   2   2   0   2   6   5

STEEL_KIND      C0   C1      C3      T0      T1   T3      T5      T7      T8
SCALE
0            0.577  0.0  0.857  0.867  0.889  1.0  0.951  0.829  0.947
1            0.423  1.0  0.143  0.133  0.111  0.0  0.049  0.171  0.053
```

: 강종에 따라 불량률에 차이가 크게 나타났지만, C0의 표본 수에 비해
나머지 강종의 표본 수가 너무 작아서 의미있는 결과를 도출해내기 힘들었다.

3 탐색적 분석 – 범주형 변수

2. FUR_NO

```
1 fun_print_crosstab(df_raw, "FUR_NO")
```

```
FUR_NO   1호기   2호기   3호기
SCALE
0         166   166   151
1          73    70    88

FUR_NO     1호기     2호기     3호기
SCALE
0         0.695   0.703   0.632
1         0.305   0.297   0.368
```

: 1, 2, 3호기 사이에 불량률에 차이가 나타났지만, 크지 않기에 유의미하다고 보기 힘들었다.

3 탐색적 분석 – 범주형 변수

3. HSB

```
1 fun_print_crosstab(df_raw, "HSB")
```

```
HSB     미적용    적용
SCALE
0          0   483
1         33   198

HSB     미적용       적용
SCALE
0       0.0   0.709
1       1.0   0.291
```

: 미적용되는 경우에는 100% 불량으로, 적용되는 경우와 큰 차이를 보이기에
주요한 설명변수라고 판단할 수 있었다.

3 탐색적 분석 – 범주형 변수

2.0 완벽히 일치하는 데이터 비교 및 확인

```
# df_raw["FUR_SZ_TEMP"] == df_raw["FUR_EXTEMP"] 인 경우가 많이 관찰되어 모든 데이터를 대상으로 테스트한 결과,

(df_raw["FUR_SZ_TEMP"] == df_raw["FUR_EXTEMP"]).value_counts()

# "FUR_SZ_TEMP"와 FUR_EXTEMP는 완벽히 일치하는 것을 알 수 있다.
```

```
True    714
dtype: int64
```

: 완벽히 일치하는 데이터를 가진 Column을 발견 → 다중공선성을 가지므로 하나의 변수를 제거.

3 탐색적 분석 – 범주형 변수

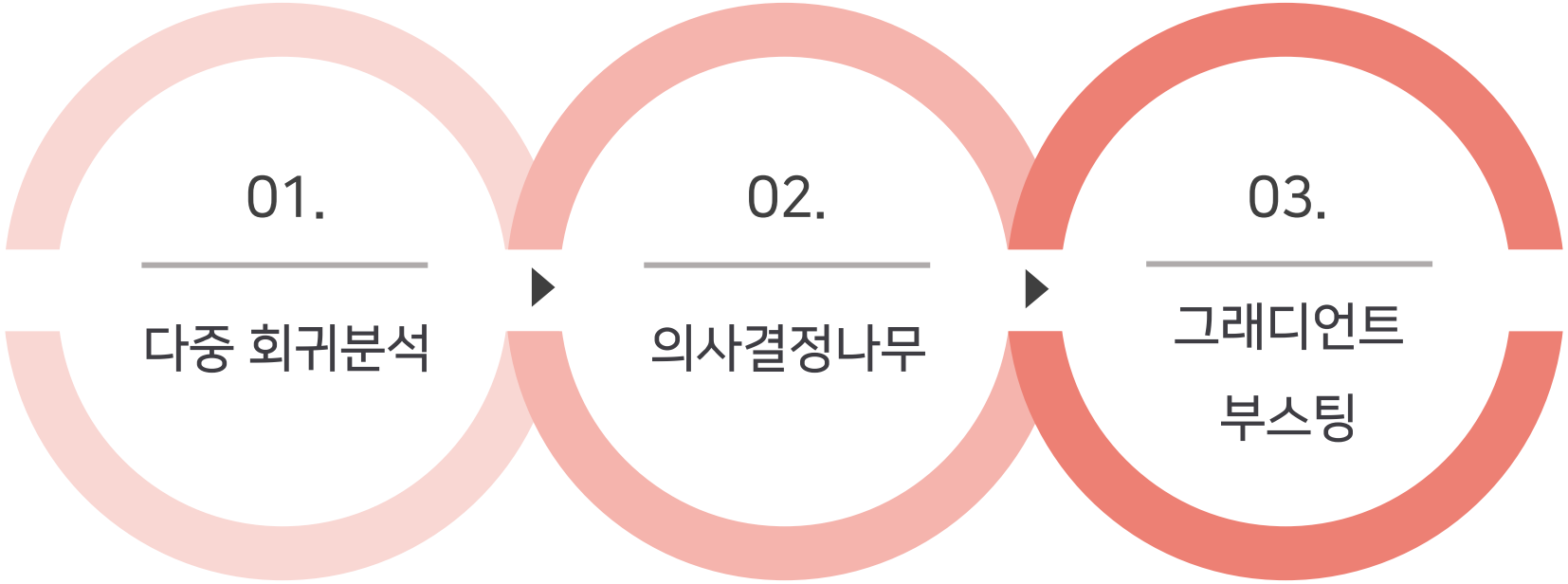
2.0 완벽히 일치하는 데이터 비교 및 확인

```
# df_raw["FUR_SZ_TEMP"] == df_raw["FUR_EXTEMP"] 인 경우가 많이 관찰되어 모든 데이터를 대상으로 테스트한 결과,

(df_raw["FUR_SZ_TEMP"] == df_raw["FUR_EXTEMP"]).value_counts()

# "FUR_SZ_TEMP"와 FUR_EXTEMP는 완벽히 일치하는 것을 알 수 있다.
```

```
True    714
dtype: int64
```

: 완벽히 일치하는 데이터를 가진 Column을 발견 → 다중공선성을 가지므로 하나의 변수를 제거.

4 모델링



4 모델링 – 01. 로지스틱 선형 회귀

Logit Regression Results

Dep. Variable:	SCALE	No. Observations:	499
Model:	Logit	Df Residuals:	476
Method:	MLE	Df Model:	22
Date:	Wed, 25 Nov 2020	Pseudo R-squ.:	0.7749
Time:	04:46:23	Log-Likelihood:	-70.461
converged:	False	LL-Null:	-313.05
Covariance Type:	nonrobust	LLR p-value:	8.970e-89

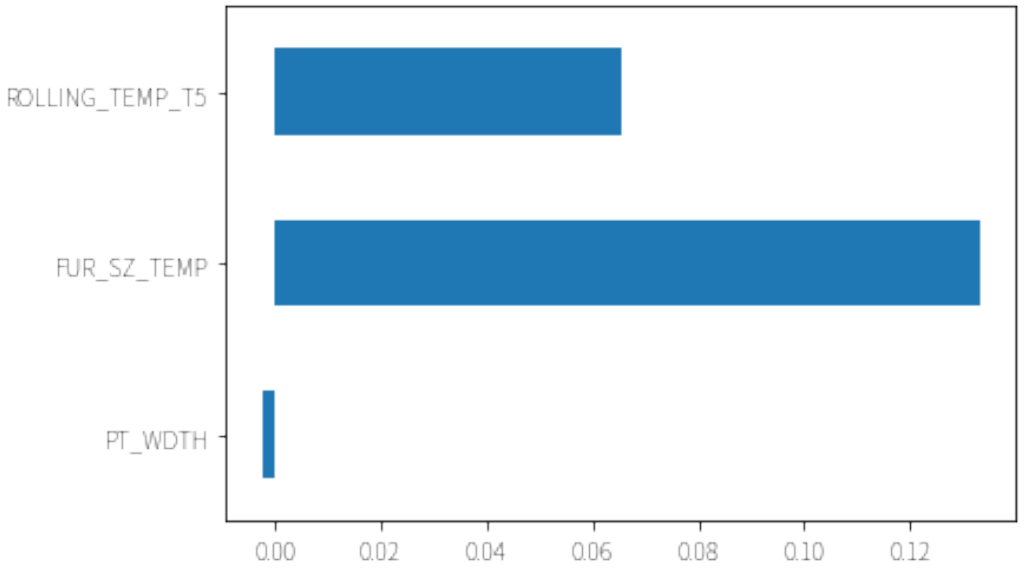
	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-164.5071	4.55e+04	-0.004	0.997	-8.93e+04	8.9e+04
C(FUR_NO)[T.2호기]	-0.8145	0.553	-1.474	0.140	-1.897	0.269
C(FUR_NO)[T.3호기]	0.2604	0.586	0.444	0.657	-0.889	1.409
C(FUR_NO_ROW)[T.2]	0.0884	0.476	0.186	0.853	-0.845	1.021
C(HSB)[T.적용]	-35.3681	2.76e+04	-0.001	0.999	-5.41e+04	5.41e+04
C(ROLLING_DESCALING)[T.6]	-42.1367	3.62e+04	-0.001	0.999	-7.09e+04	7.08e+04
C(ROLLING_DESCALING)[T.7]	-1.4370	8.2e+06	-1.75e-07	1.000	-1.61e+07	1.61e+07
C(ROLLING_DESCALING)[T.8]	-38.3731	3.62e+04	-0.001	0.999	-7.09e+04	7.08e+04
C(ROLLING_DESCALING)[T.9]	1.8885	1.89e+06	9.97e-07	1.000	-3.71e+06	3.71e+06
C(ROLLING_DESCALING)[T.10]	-38.5217	3.62e+04	-0.001	0.999	-7.09e+04	7.08e+04
C(WORK_GR)[T.2조]	-0.7646	0.627	-1.219	0.223	-1.994	0.465
C(WORK_GR)[T.3조]	-1.1117	0.708	-1.570	0.117	-2.500	0.276
C(WORK_GR)[T.4조]	-0.8144	0.636	-1.280	0.200	-2.061	0.432
PT_THK	0.1521	0.093	1.632	0.103	-0.031	0.335
PT_WDTH	-0.0024	0.001	-3.030	0.002	-0.004	-0.001
PT_LTH	-5.69e-05	4.52e-05	-1.258	0.208	-0.000	3.17e-05
PT_WGT	-1.268e-05	1.05e-05	-1.205	0.228	-3.33e-05	7.95e-06
FUR_HZ_TEMP	0.0241	0.032	0.761	0.446	-0.038	0.086
FUR_HZ_TIME	0.0089	0.007	1.227	0.220	-0.005	0.023
FUR_SZ_TEMP	0.1332	0.065	2.051	0.040	0.006	0.260
FUR_SZ_TIME	-0.0302	0.020	-1.522	0.128	-0.069	0.009
FUR_TIME	-0.0038	0.007	-0.538	0.590	-0.018	0.010
ROLLING_TEMP_T5	0.0653	0.012	5.612	0.000	0.043	0.088

다중 회귀분석 결과,

1 위 결과를 분석해보면, p-value를 고려해보았을 때, "PT_WDTH", "FUR_SZ_TEMP", "ROLLING_TEMP_T5" 의 3가지 요인이 SCALE 발생여부에 영향을 주는 것으로 판단할 수 있다.

2 PT_WDTH와는 음의 상관관계, 그리고 FUR_SZ_TEMP, ROLLING_TEMP_T5와는 양의 상관관계가 나타난다.

4 모델링 – 01. 로지스틱 선형 회귀



Coefficient 값을 구해보았을 때,
유효한 값은 다음의 세 개 뿐이었다.

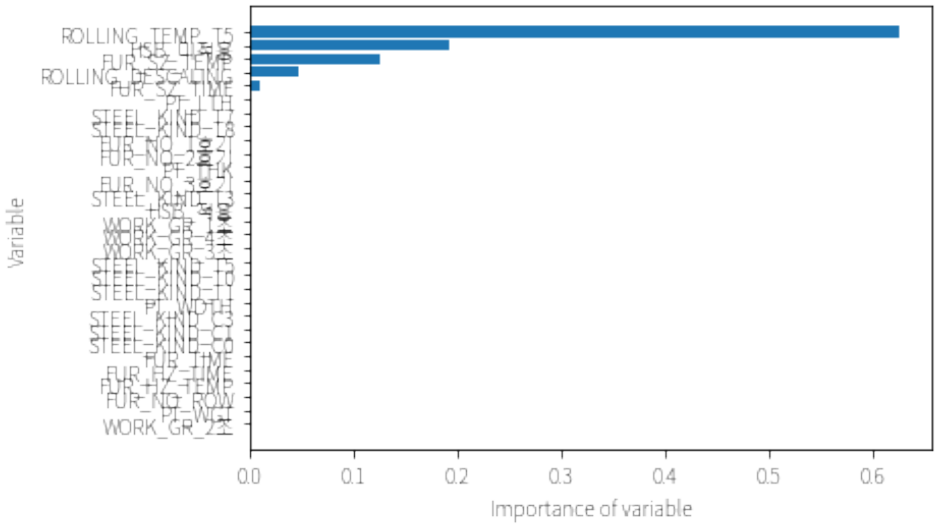
: 로지스틱 선형회귀를 하였을 때, FUR_SZ_TEMP, ROLLING_TEMP_T5, PT_WIDTH 순서로 영향을 미치는 것을 확인할 수 있었다.

: 우리가 제조공정에서 변화를 줄 수 있는 변수는 FUR_SZ_TEMP, ROLLING_TEMP_T5로, 이 두가지가 Vital Few(VF) 라고 판단할 수 있다.

4 모델링 – 02. 의사결정나무

```
Score on training set: 0.978
Score on test set: 0.953
Confusion matrix:
[[144   0]
 [ 10  61]]
```

	Feature	Importance
10	ROLLING_TEMP_T5	0.626
24	HSB_미적용	0.193
7	FUR_SZ_TEMP	0.125
11	ROLLING_DESCALING	0.047
8	FUR_SZ_TIME	0.010
19	STEEL_KIND_T7	0.000
20	STEEL_KIND_T8	0.000
21	FUR_NO_1호기	0.000
22	FUR_NO_2호기	0.000



1. Training Set : 0.978(97.8%) / Test Set : 0.953(95.3%)
2. Importance (트리를 만들 때의) : FUR_SZ_TEMP < HSB < ROLLING_TEMP_T5

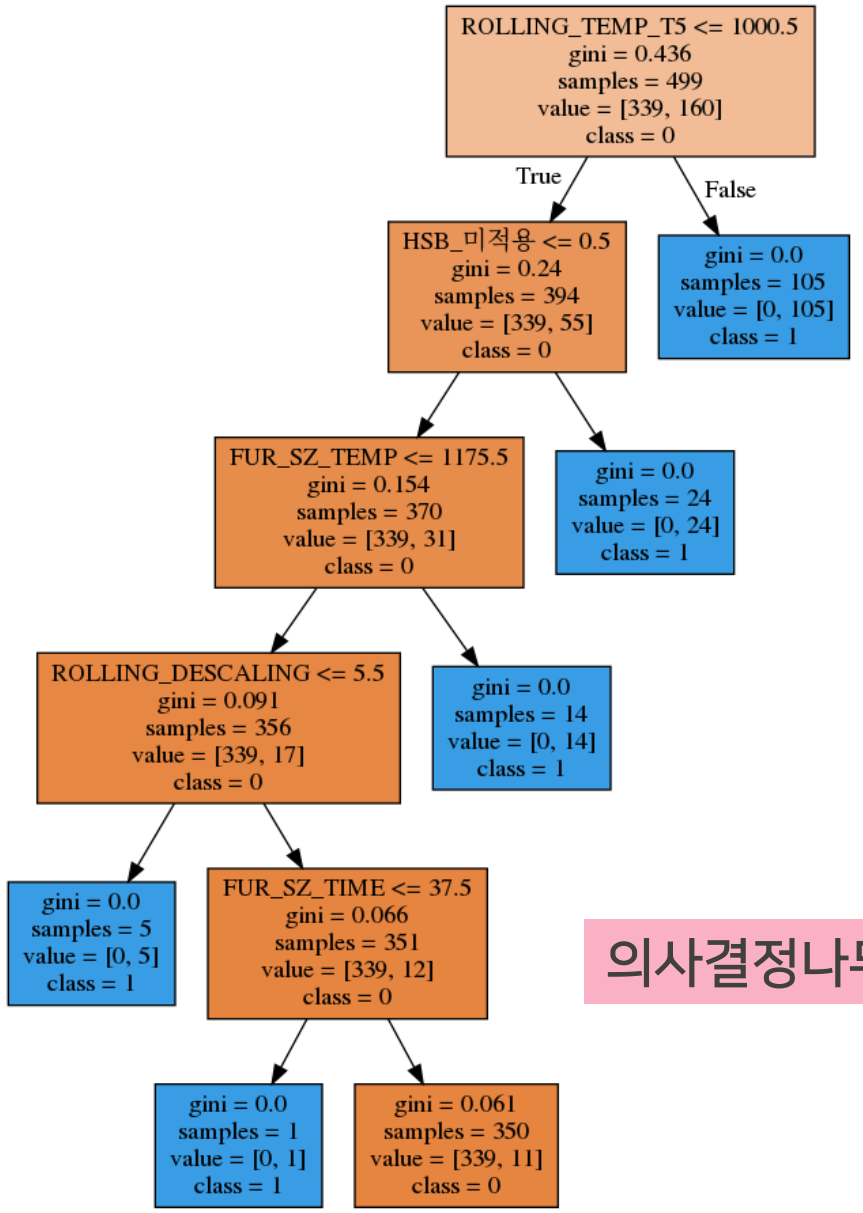
4 모델링 – 02. 의사결정나무

1 "ROLLING_TEMP_T5 " < 1000.5

2 HSB_미적용 = 0

3 FUR_SZ_TEMP <= 1175.5

→ 위 조건들을 통과해야 한다.

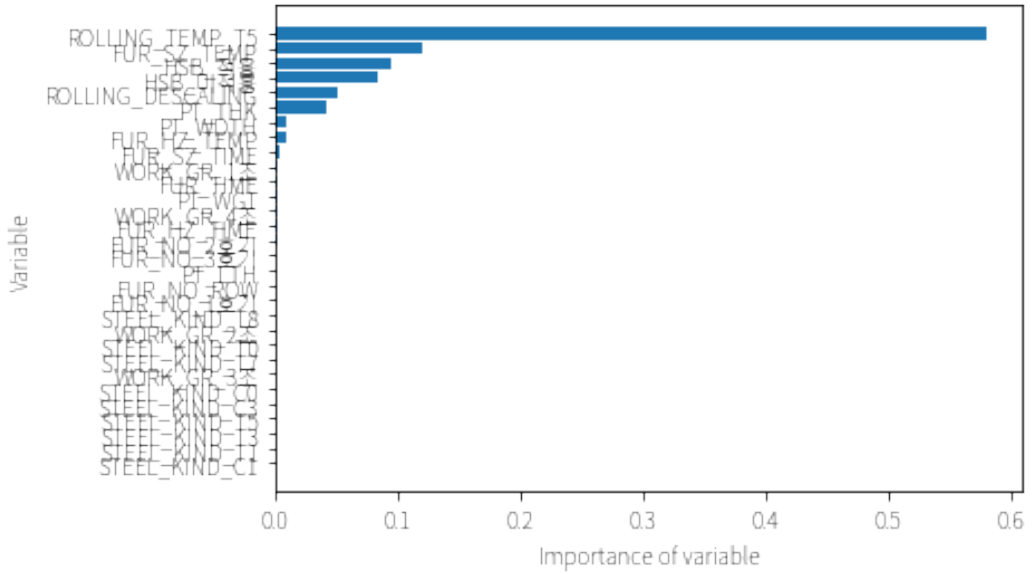


의사결정나무

4 모델링 – 03. 그래디언트 부스팅

```
Accuracy on training set: 1.000
Accuracy on test set: 0.972
Confusion matrix:
[[144   0]
 [  6  65]]
```

	Feature	Importance
10	ROLLING_TEMP_T5	0.580
7	FUR_SZ_TEMP	0.120
25	HSB_적용	0.095
24	HSB_미적용	0.084
11	ROLLING_DESCALING	0.051



1. Training Set : 1.000(100.0%) / Test Set : 0.972(97.2%)
2. Importance (트리를 만들 때의) : HSB < FUR_SZ_TEMP < ROLLING_TEMP_T5

6 소감

데이터 분석을 하며,

데이터에서 이상치를 이상치라 판단하고, 데이터에서 통찰을 얻는 것은
Domain 지식을 바탕으로 얻을 수 있는 역량임을 더 깊이 깨달았습니다.

데이터 분석 역량 뿐 아니라, 현재 가지고 있는 기계공학 전공 지식 역량을
꾸준히 발전시켜야겠다고 다시 한번 다짐하게 되었습니다.