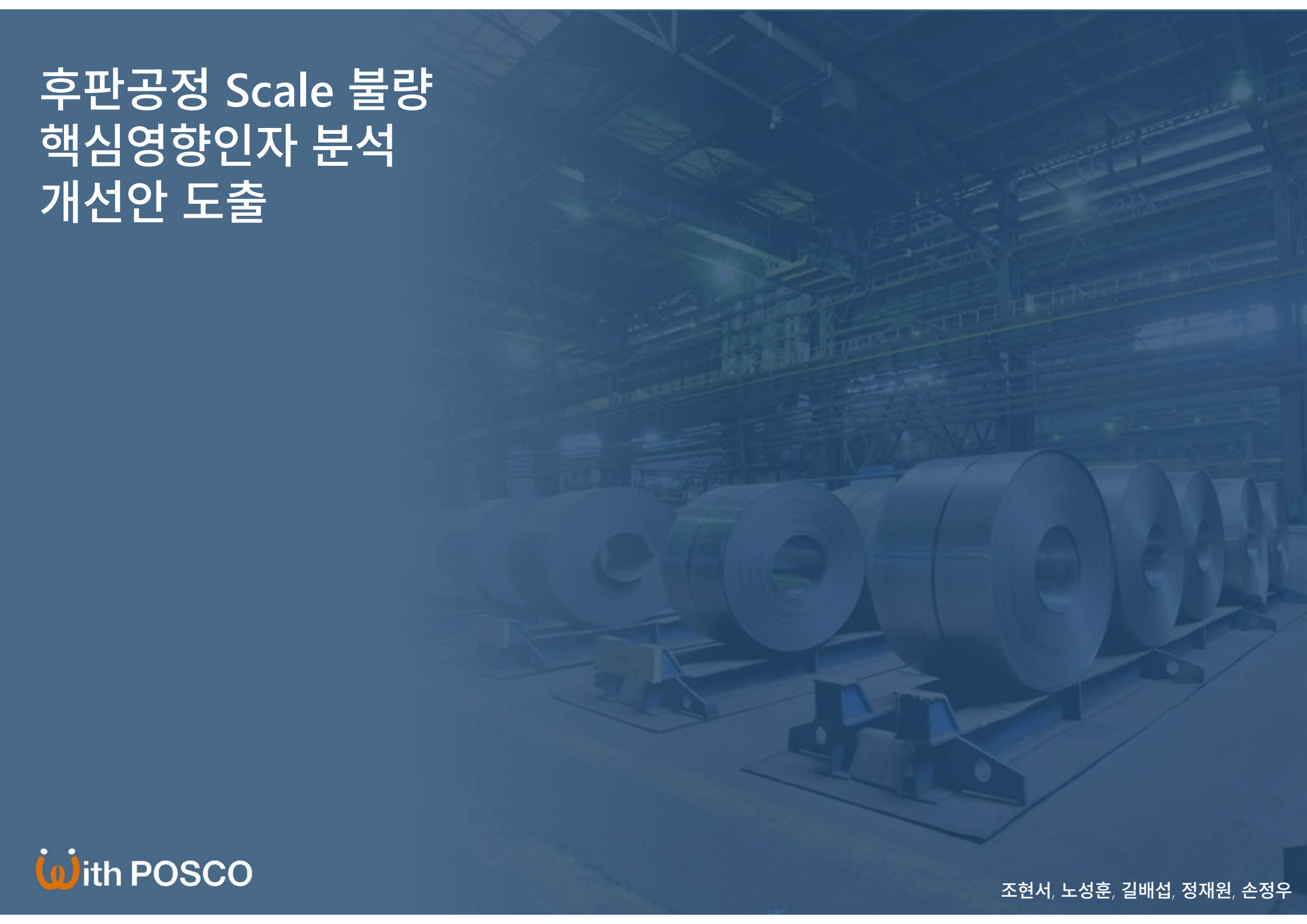
후판공정 Scale 불량 핵심영향인자 분석 개선안 도출



With POSCO

조현서, 노성훈, 길배섭, 정재원, 손정우

과제 정의

주제	• 급증하는 압연공정 Scale 불량의 원인 분석 및 개선 방안 도출

과제 정의

1. Pos Scale 불량 데이터에 기술통계 적용 계획 수립
2. Scale 불량 데이터의 속성을 특정하여 통계량을 사용해 정리, 요약, 설명
3. Scale 불량에 핵심영향인자가 무엇인지 파악
4. Scale 불량을 예방하고 현장 개선안 도출
5. 방안 도출을 위해서 과정 및 피드백 기술

목표 정의

후판 제품 Scale 불량 원인 분석 및 불량 발생률 0% 달성을 위한

핵심영향인자 도출과 모델 개선 방안 도출

분석계획

1. 데이터 현황
2. 데이터 전처리
3. 가설검정 및 그래프 탐색
4. 모델링
5. 핵심영향인자 도출
6. 개선안 제시

With POSCO

조현서, 노성훈, 길배섭, 정재원, 손정우

분석 계획

- **다음과 같은 절차에 따라 EDA(탐색적 데이터 분석) 실시**

데이터셋 정제

원시 데이터 수집

- 후판공정 센서 데이터 수 ('SCALE불량.csv')

데이터 전처리

- 결측치 처리
- 이상치 확인
  - 연속형 변수: boxplot 활용
  - 범주형 변수: 개별 값 확인

변수 파생·변형

- 변수 간 상관성 검토
- 도메인 지식 적용
- 날짜 변수 변형
  예) 연/월/일/시간 정보가 한 열에 같이 있을 경우 분리하여 유의성 확인

모델링·알고리즘

의사결정 트리 모델링

랜덤포레스트 모델링

그래디언트 부스트 모델링

XG Boost 모델링

서포트 벡터 모델 (SVM)

인공신경망 모델 (NN)

- 파라미터 조정을 통한 성능 개선
- 모델 성능 비교

Data Solution

모델 성능 개선

- Train-test 점수 기반 정확도·과적합 판단
- 최적의 하이퍼 파라미터 지정

핵심 영향인자 도출

- 최종 모델링 후 변수 중요도 도출
- 핵심 영향 요인 식별 및 분석

개선안 제시

- Scale 불량률 0% 예방 및 품질 향상 (안) 제시
- 예) 적정온도 유지, 탈탄 및 표면 품질 저하 방지 등

1. 데이터 현황

- 목표변수 ‘scale’를 예측하기 위한 데이터 정제 (자료수: 1,000, 특징: 21개)

파일 불러온 후 ‘.head()’ 실행 결과:

- Column 총 21개
- 총 21개의 변수로 구성된 데이터 (설명변수 20개)

df

	plate_no	rolling_date	scale	spec_long	spec_country	steel_kind	pt_thick	pt_width	pt_length	hsb	...	fur_input_row	fur_heat_temp	fur_heat
0	PLT_1001	03JAN2023:07:07:53	양품	AB/EH32-TM	미국	T	32	3700	15100	적용	...	1열	1144	
1	PLT_1002	03JAN2023:07:21:22	양품	AB/EH32-TM	미국	T	32	3700	15100	적용	...	2열	1144	
2	PLT_1003	03JAN2023:07:31:15	양품	NV-E36-TM	영국	T	33	3600	19200	적용	...	1열	1129	
3	PLT_1004	03JAN2023:07:41:01	양품	NV-E36-TM	영국	T	33	3600	19200	적용	...	2열	1152	
4	PLT_1005	03JAN2023:07:52:40	양품	BV-EH36-TM	프랑스	T	38	3100	13300	적용	...	1열	1140	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
995	PLT_1996	10JAN2023:05:32:25	양품	BV-A	프랑스	C	19	3400	41500	적용	...	2열	1142	
996	PLT_1997	10JAN2023:05:39:19	양품	LR-A	영국	C	19	3400	41500	적용	...	2열	1142	
997	PLT_1998	10JAN2023:05:52:41	양품	AB/AH32	미국	C	17	3400	43700	적용	...	2열	1169	
998	PLT_1999	10JAN2023:06:01:50	양품	NV-A32	영국	C	17	3400	43700	적용	...	2열	1169	
999	PLT_2000	10JAN2023:06:16:27	양품	GL-A32	독일	C	16	3400	54200	적용	...	1열	1186	

1000 rows × 21 columns

변수	변수 역할	형태
plate_no	ID	범주형
rolling_date	날짜	연속형
scale	목표변수	범주형
spec_long	설명변수	범주형
spec_country	설명변수	범주형
steel_kind	설명변수	범주형
pt_thick	설명변수	연속형
pt_width	설명변수	연속형
pt_length	설명변수	연속형
hsb	설명변수	범주형
fur_no	설명변수	범주형
fur_input_row	설명변수	범주형
fur_heat_temp	설명변수	연속형
fur_heat_time	설명변수	연속형
fur_soak_temp	설명변수	연속형
fur_soak_time	설명변수	연속형
fur_total_time	설명변수	연속형
rolling_method	설명변수	범주형
rolling_temp	설명변수	연속형
descaling_count	설명변수	연속형
work_group	설명변수	범주형

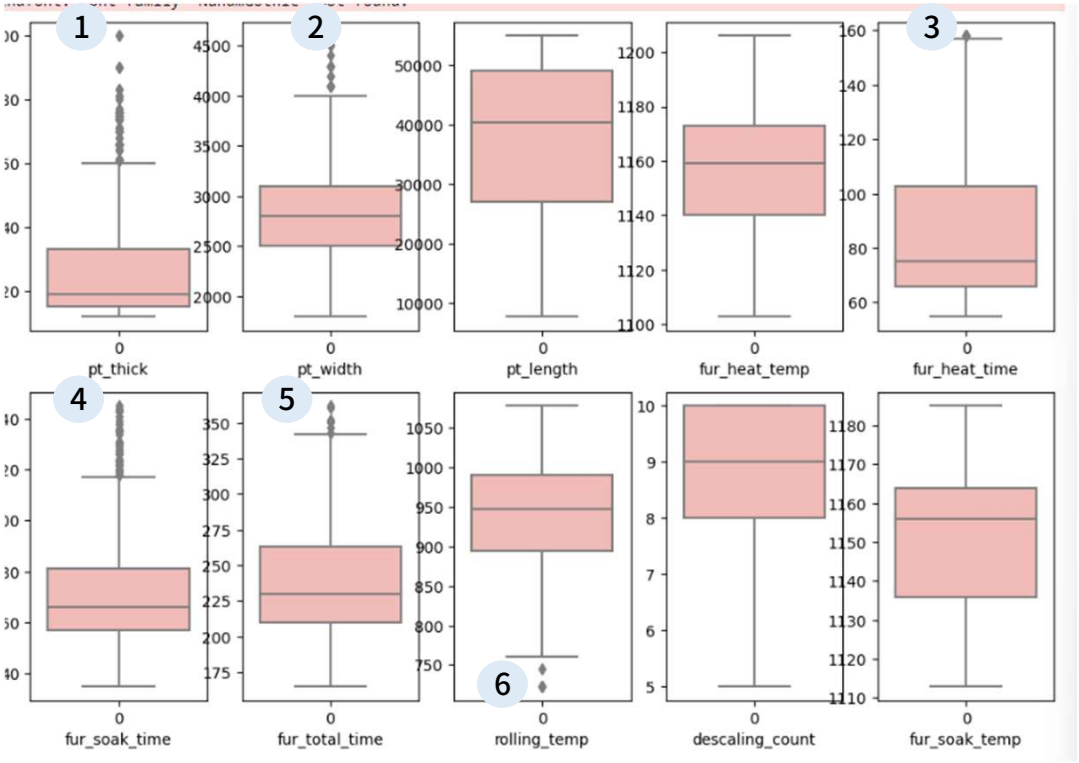
- 원본 데이터의 결측치·이상치 검토

'.info()' 실행 결과 (결측치 검토):

```
df.info();df.shape   # 연속형 10개 / 그 외 11개 범주형
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 21 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   plate_no         1000 non-null   object
 1   rolling_date     1000 non-null   object
 2   scale            1000 non-null   object
 3   spec_long        1000 non-null   object
 4   spec_country     1000 non-null   object
 5   steel_kind       1000 non-null   object
 6   pt_thick         1000 non-null   int64
 7   pt_width         1000 non-null   int64
 8   pt_length        1000 non-null   int64
 9   hsb              1000 non-null   object
 10  fur_no           1000 non-null   object
 11  fur_input_row    1000 non-null   object
 12  fur_heat_temp    1000 non-null   int64
 13  fur_heat_time    1000 non-null   int64
 14  fur_soak_temp    1000 non-null   int64
 15  fur_soak_time    1000 non-null   int64
 16  fur_total_time   1000 non-null   int64
 17  rolling_method   1000 non-null   object
 18  rolling_temp     1000 non-null   int64
 19  descaling_count  1000 non-null   int64
 20  work_group       1000 non-null   object
dtypes: int64(10), object(11)
memory usage: 164.2+ KB

(1000, 21)
```

.Boxplot 활용 결과 (이상치 확인):



이상치 처리			근거
Pt_thick	>= 81.17	유지	이상치 중 수치가 제일 높은 100의 값 2개를 제거하였습니다
Pt_width	>= 4314.2	유지	이상치가 7개 존재하지만, 평균과 표준편차를 비교한 결과 큰 문제가 없다고 판단하여 처리하지 않았습니다.
Fur_heat_time	>= 164.8	제거	이상치가 존재하지 않습니다.
Fur_soak_time	>= 133.5	유지	이상치가 존재하지만, 평균과 편차를 고려했을 때 제거하지 않았습니다.
Fur_total_time	>= 352.8	유지	이상치가 존재하지만, 평균과 편차를 고려했을 때 제거하지 않았습니다.
Rolling_temp	<=748.6	4개 제거	이상치가 존재하지만, 평균과 표준편차를 고려 했을때 제거하지 않았습니다
Rolling_temp	== 0	4개 제거	수치가 0인 값이 6개를 발견하고 제거하였습니다.

- 변수 파생 및 변형

```
df.info();df.shape    # 연속형 10개 / 그 외 11개 범주형
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 21 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   plate_no         1000 non-null   object
 1   rolling_date     1000 non-null   object
 2   scale            1000 non-null   object
 3   spec_long        1000 non-null   object
 4   spec_country     1000 non-null   object
 5   steel_kind       1000 non-null   object
 6   pt_thick         1000 non-null   int64
 7   pt_width         1000 non-null   int64
 8   pt_length        1000 non-null   int64
 9   hsb              1000 non-null   object
 10  fur_no           1000 non-null   object
 11  fur_input_row    1000 non-null   object
 12  fur_heat_temp    1000 non-null   int64
 13  fur_heat_time    1000 non-null   int64
 14  fur_soak_temp    1000 non-null   int64
 15  fur_soak_time    1000 non-null   int64
 16  fur_total_time   1000 non-null   int64
 17  rolling_method   1000 non-null   object
 18  rolling_temp     1000 non-null   int64
 19  descaling_count  1000 non-null   int64
 20  work_group       1000 non-null   object
dtypes: int64(10), object(11)
memory usage: 164.2+ KB

(1000, 21)
```

범주형 설명변수의 '.unique()' 수행 결과 별도 이상치 없었음.

['scale' 원 핫 인코딩]

```
df["scale"].value_counts() #  Scale(산화철) 불량 목표변수 정상 690개 불량 310개

양품    690
불량    310
Name: scale, dtype: int64
```

scale "양품" = 0, scale "불량" = 1

[Rolling 일자 & 시간 변수 파생]

hsb	fur_no	fur_input_row	...	fur_heat_time	fur_soak_temp	fur_soak_time	fur_total_time	rolling_method	rolling_temp	descaling_count	work_group	day	hour
적용	1	1	...	116	1133	59	259	TMCP(온도제어)	934	8		03	07
적용	1	2	...	122	1135	53	238	TMCP(온도제어)	937	8		03	07
적용	2	1	...	116	1121	55	258	TMCP(온도제어)	889	8		03	07
적용	2	2	...	125	1127	68	266	TMCP(온도제어)	885	8		03	07
적용	3	1	...	134	1128	48	246	TMCP(온도제어)	873	8		03	07
...	...	...	...	...	...	...	...	...	...	...	...	...	...

'rolling_date' → 'day', 'date', 'time', 'hour'

[가열로 호기·열 변수 파생]

ur_input_row	fur_heat_temp	...	fur_soak_temp	fur_soak_time	fur_total_time	rolling_method	rolling_temp	descaling_count	work_group	day	hour	fur_combined
1	1144	...	1133	59	259	TMCP(온도제어)	934	8	1	03	07	1_1
2	1144	...	1135	53	238	TMCP(온도제어)	937	8	1	03	07	1_2
1	1129	...	1121	55	258	TMCP(온도제어)	889	8	1	03	07	2_1
2	1152	...	1127	68	266	TMCP(온도제어)	885	8	1	03	07	2_2
1	1140	...	1128	48	246	TMCP(온도제어)	873	8	1	03	07	3_1
...	...	...	...	...	...	...	...	...	...	...	...	...

'fur_no' + 'fur_input_row' → 'fur_combined'

- **데이터 전처리 전/후 비교**

원본 데이터 분포

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 21 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   plate_no         1000 non-null   object
 1   rolling_date     1000 non-null   object
 2   scale            1000 non-null   object
 3   spec_long        1000 non-null   object
 4   spec_country     1000 non-null   object
 5   steel_kind       1000 non-null   object
 6   pt_thick         1000 non-null   int64
 7   pt_width         1000 non-null   int64
 8   pt_length        1000 non-null   int64
 9   hsb              1000 non-null   object
 10  fur_no           1000 non-null   object
 11  fur_input_row    1000 non-null   object
 12  fur_heat_temp    1000 non-null   int64
 13  fur_heat_time    1000 non-null   int64
 14  fur_soak_temp    1000 non-null   int64
 15  fur_soak_time    1000 non-null   int64
 16  fur_total_time   1000 non-null   int64
 17  rolling_method   1000 non-null   object
 18  rolling_temp     1000 non-null   int64
 19  descaling_count  1000 non-null   int64
 20  work_group       1000 non-null   object
dtypes: int64(10), object(11)
memory usage: 164.2+ KB
```

데이터 전처리 후 데이터 분포

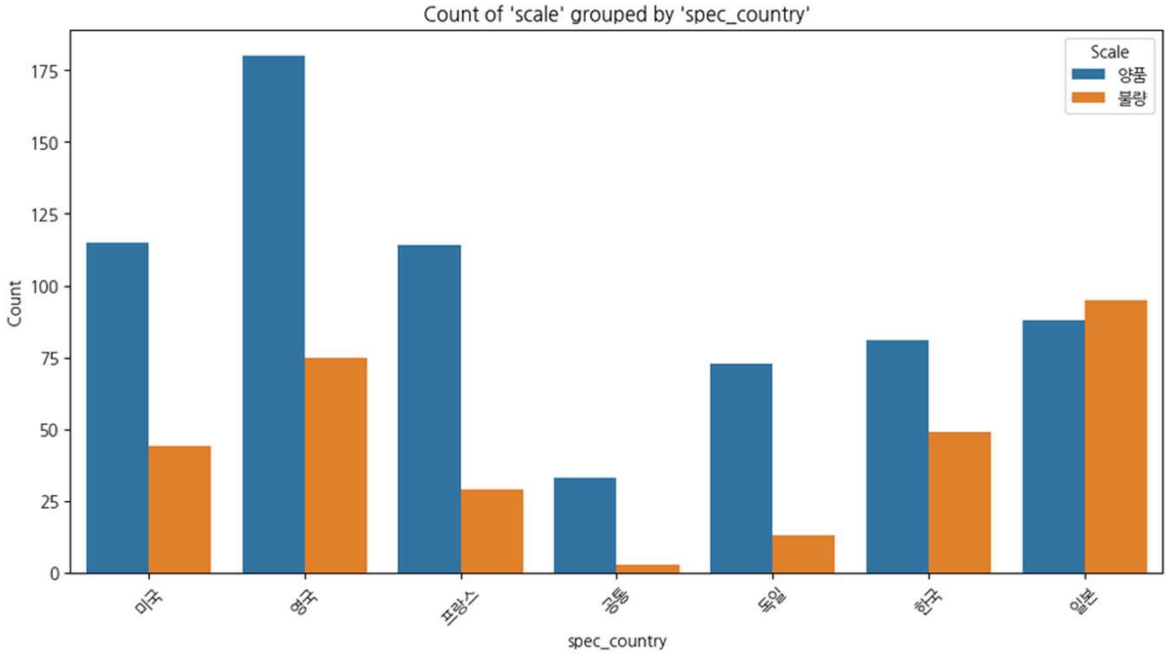
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 992 entries, 0 to 999
Data columns (total 23 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   spec_long        992 non-null    object
 1   spec_country     992 non-null    object
 2   steel_kind       992 non-null    object
 3   pt_thick         992 non-null    int64
 4   pt_width         992 non-null    int64
 5   pt_length        992 non-null    int64
 6   hsb              992 non-null    object
 7   fur_no           992 non-null    object
 8   fur_input_row    992 non-null    object
 9   fur_heat_temp    992 non-null    int64
 10  fur_heat_time    992 non-null    int64
 11  fur_soak_temp    992 non-null    int64
 12  fur_soak_time    992 non-null    int64
 13  fur_total_time   992 non-null    int64
 14  rolling_method   992 non-null    object
 15  rolling_temp     992 non-null    int64
 16  descaling_count  992 non-null    int64
 17  work_group       992 non-null    object
 18  fur_combined     992 non-null    category
 19  day              992 non-null    object
 20  hour             992 non-null    object
 21  scale_불량         992 non-null    uint8
 22  scale            992 non-null    uint8
dtypes: category(1), int64(10), object(10), uint8(2)
memory usage: 198.2+ KB
```

- //

탐색적 분석 _(5) 변수별 검정

- Spec_country와 scale의 독립성 가설검정 → 제품 규격별 scale 차이 有

Spec_country




```
chi-squared statistic: 67.67587628130671
p-value: 1.2240601602506334e-12
degrees of freedom: 6
expected frequencies:
 [[ 11.17741935  26.7016129   49.36693548  79.1733871   56.81854839
   44.39919355  40.36290323]
 [ 24.82258065  59.2983871  109.63306452 175.8266129  126.18145161
   98.60080645  89.63709677]]
```

귀무가설	(H0) : ‘spec_country’와 ‘scale’ 변수는 독립이다.
대립가설	(H1) : ‘spec_country’와 ‘scale’ 변수는 독립적이지 않다.

[결론] : 대립가설 채택

- “제품 규격 과 불량률은 관련이 있다.”

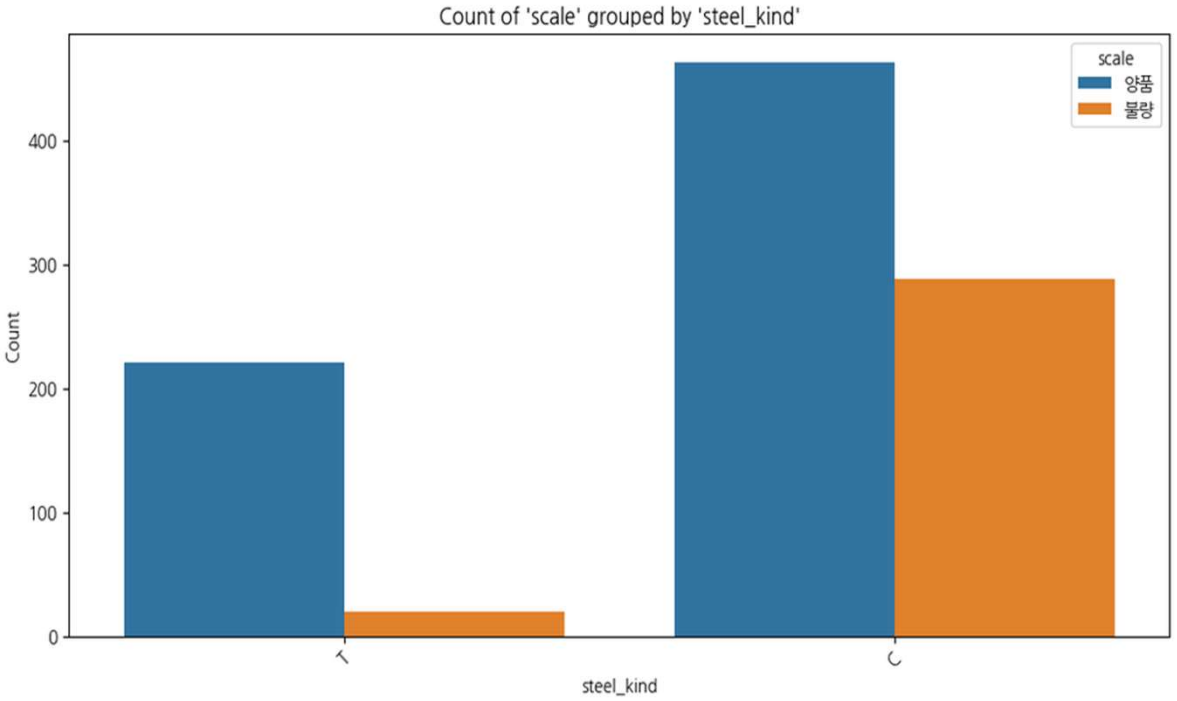
[근거]

- P-값이 0에 수렴한다.
- 유의수준 5%에서 귀무가설을 기각할 수 있다.
- 그러므로, 제품 규격과 불량률은 관련이 있다.

탐색적 분석 _(5) 변수별 검정

- Steel_kind와 scale의 독립성 가설검정 → 강종별 scale 차이 有

Steel_kind




```
chi-squared statistic: 75.56078768326427
p-value: 3.543287859042724e-18
degrees of freedom: 1
expected frequencies:
 [[233.1733871  74.8266129]
 [517.8266129 166.1733871]]
```

귀무가설	(H0): ‘scale’변수와 ‘steel_kind’ 변수 독립이다.
대립가설	(H1): ‘scale’변수와 ‘steel_kind’ 변수는 독립이지 않다.

[결론] : 대립가설 채택

•“강종과 불량률은 관련이 있다.”

[근거]

•P-값이 0에 수렴한다.

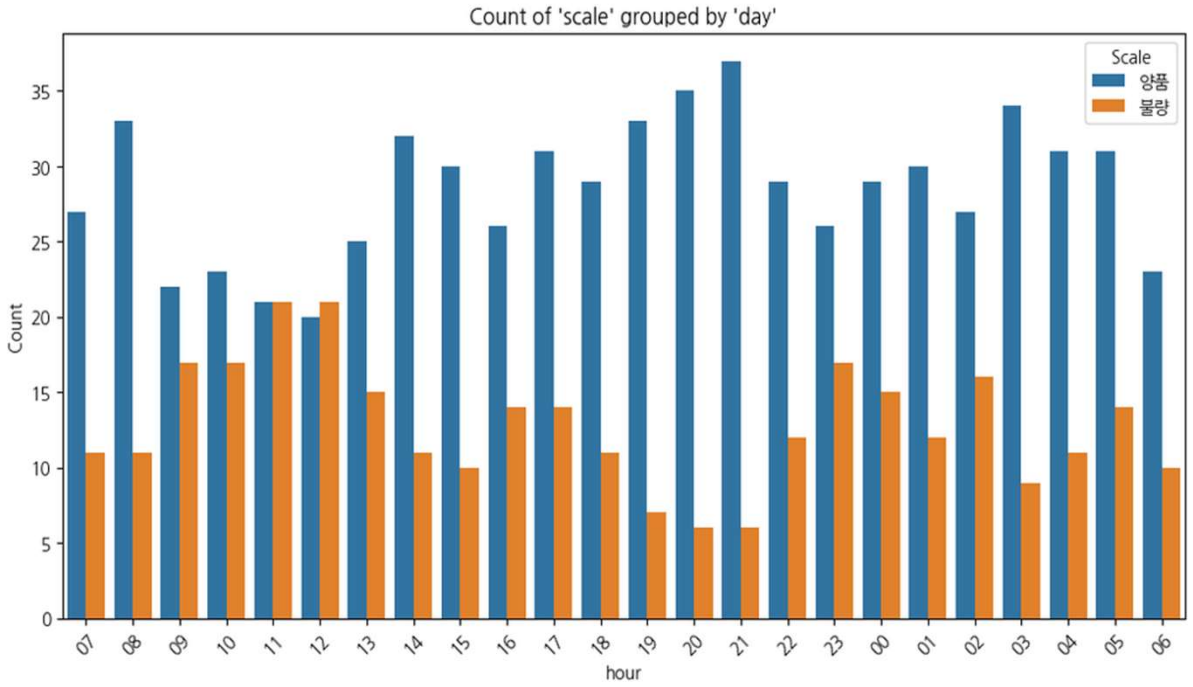
•유의수준 5%에서 귀무가설을 기각할 수 있다.

•그러므로, 강종과 불량률은 관련이 있다.

탐색적 분석 _(5) 변수별 검정

- Hour와 scale의 독립성 가설검정 → 시간별 scale 차이 有

Hour




```
chi-squared statistic: 43.14039558813147
p-value: 0.006664858637988756
degrees of freedom: 23
expected frequencies:
```

귀무가설	(H0): ‘hour’ 변수와 ‘scale’ 변수는 독립이다.
대립가설	(H1): ‘hour’ 변수와 ‘scale’ 변수는 독립적이지 않다.

[결론] : 대립가설 채택

- “시간과 불량률은 관련이 있다.”

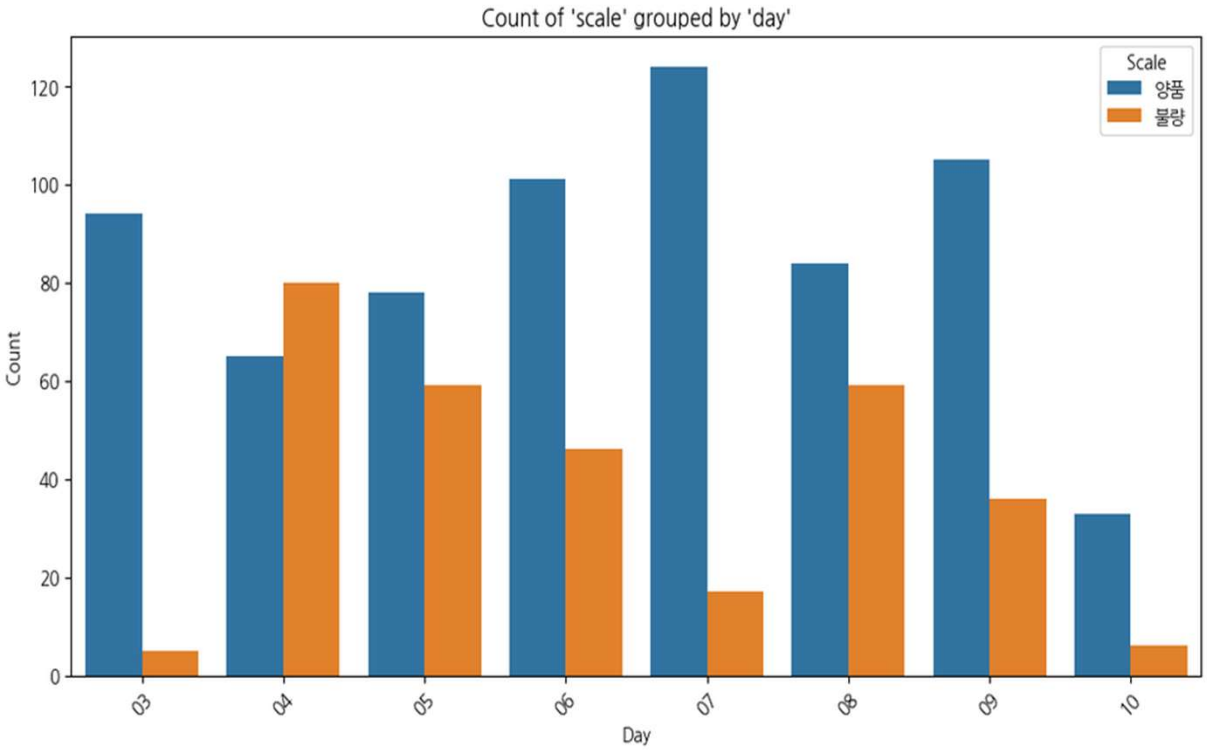
[근거]

- P-값이 0에 수렴한다.
- 유의수준 5%에서 귀무가설을 기각할 수 있다.
- 그러므로, 시간과 불량률은 관련이 있다.

탐색적 분석 _(5) 변수별 검정

- Day와 scale의 독립성 가설검정 → 날짜별 scale 차이 有

Day




```
chi-squared statistic: 117.11128587001863
p-value: 3.0589281000689563e-22
degrees of freedom: 7
```

귀무가설	(H0): ‘day’ 변수와 ‘scale’ 변수는 독립이다.
대립가설	(H1): ‘day’ 변수와 ‘scale’ 변수는 독립적이지 않다.

[결론] : 대립가설 채택

- “날짜와 불량률은 관련이 있다.”

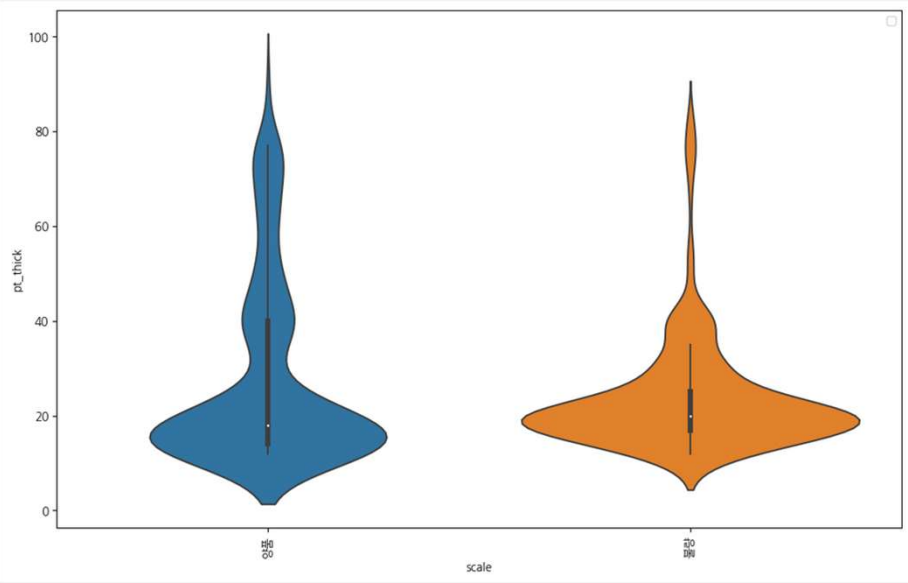
[근거]

- P-값이 0에 수렴한다.
- 유의수준 5%에서 귀무가설을 기각할 수 있다.
- 그러므로, 날짜와 불량률은 관련이 있다.

탐색적 분석 _(5) 변수별 검정

- Pt_thick와 scale의 독립성 가설검정 → 후판 지시 두께별 scale 차이 有

Pt_thick




```
Optimization terminated successfully.
         Current function value: 0.613191
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:                  scale   No. Observations:                  992
Model:                          Logit   Df Residuals:                      990
Method:                           MLE   Df Model:                            1
Date:                Tue, 28 May 2024   Pseudo R-squ.:                0.01016
Time:                        19:43:31   Log-Likelihood:                -608.29
converged:                       True   LL-Null:                       -614.53
Covariance Type:            nonrobust   LLR p-value:                 0.0004089
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.4220      0.127      3.316      0.001       0.173       0.671
pt_thick       0.0147      0.004      3.382      0.001       0.006       0.023
==============================================================================
```

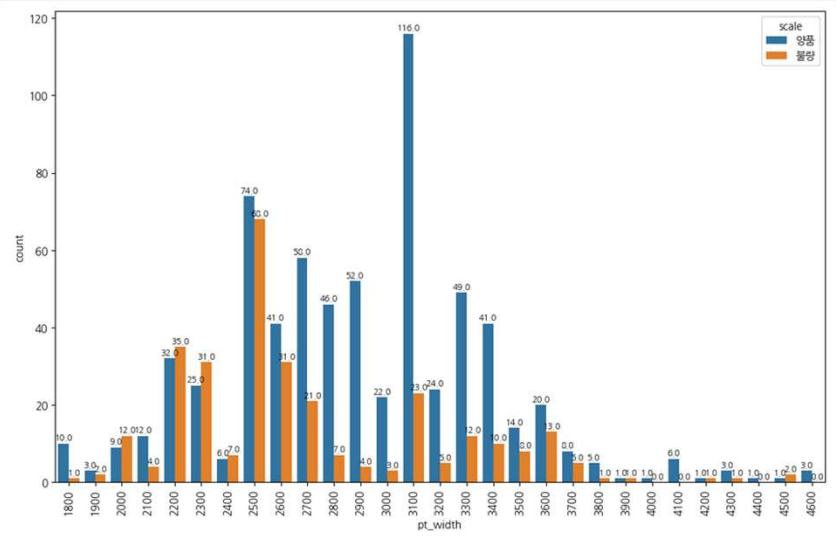
귀무가설	(H0): ‘pt_thick’ 변수와 ‘scale’ 변수는 독립이다.
대립가설	(H1): ‘pt_thick’ 변수와 ‘scale’ 변수는 독립적이지 않다.

- P-값이 0.05보다 작다.
- 유의수준 5%에서 귀무가설을 기각할 수 있다
- **[결론] 후판 지시 두께와 불량률은 관련이 있다.**

탐색적 분석 _(5) 변수별 검정

- Pt_width와 scale의 독립성 가설검정 → 후판 지시 폭별 scale 차이 有

Pt_width




```
Optimization terminated successfully.
         Current function value: 0.600382
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:                  scale   No. Observations:                  992
Model:                          Logit   Df Residuals:                      990
Method:                           MLE   Df Model:                            1
Date:                Tue, 28 May 2024   Pseudo R-squ.:                0.03084
Time:                        19:43:32   Log-Likelihood:                -595.58
converged:                       True   LL-Null:                       -614.53
Covariance Type:            nonrobust   LLR p-value:                 7.431e-10
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.7313      0.427     -4.058      0.000      -2.567      -0.895
pt_width       0.0009      0.000      5.923      0.000       0.001       0.001
==============================================================================
```

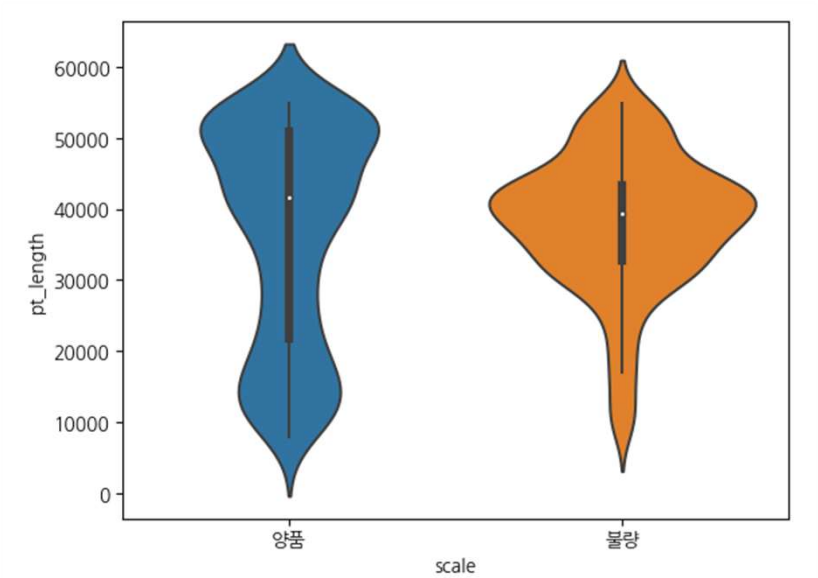
귀무가설	(H0): ‘pt_width’ 변수와 ‘scale’ 변수는 독립이다.
대립가설	(H0): ‘pt_width’ 변수와 ‘scale’ 변수는 독립이다.

- P-값이 0.05보다 작다.
- 유의수준 5%에서 귀무가설을 기각할 수 있다.
- **[결론] “후판 지시폭은 불량률은 관련이 있다.**

탐색적 분석 _(5) 변수별 검정

- Pt_length와 scale의 독립성 가설검정 → 후판 지시 길이별 scale 차이 無

Pt_length




```
Optimization terminated successfully.
         Current function value: 0.617950
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:                  scale   No. Observations:                  992
Model:                          Logit   Df Residuals:                      990
Method:                           MLE   Df Model:                            1
Date:                Tue, 28 May 2024   Pseudo R-squ.:                0.002482
Time:                        19:43:32   Log-Likelihood:                -613.01
converged:                       True   LL-Null:                       -614.53
Covariance Type:            nonrobust   LLR p-value:                   0.08074
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.1239      0.202      5.574      0.000       0.729       1.519
pt_length  -8.756e-06   5.05e-06     -1.735      0.083   -1.86e-05    1.13e-06
==============================================================================
```

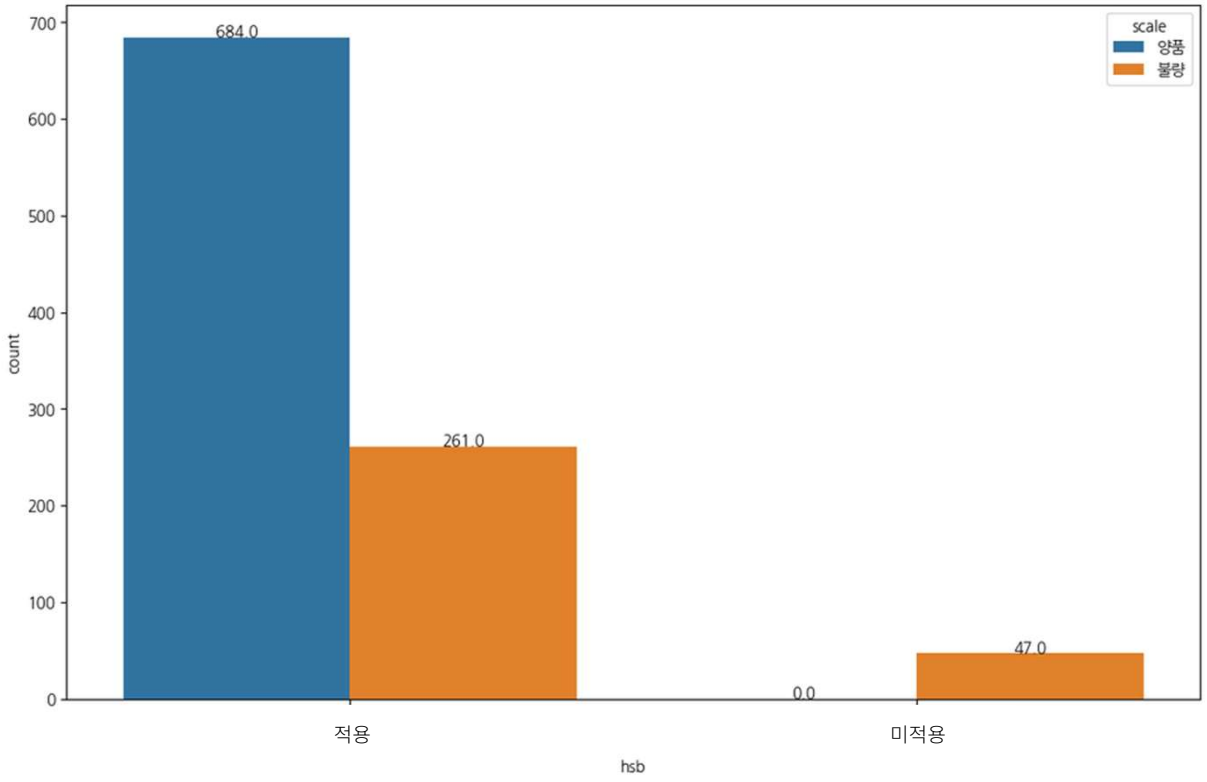
귀무가설	(H0): ‘pt_length’ 변수와 ‘scale’ 변수는 독립이다.
대립가설	(H1): ‘pt_length’ 변수와 ‘scale’ 변수는 독립적이지 않다.

- P-값이 0.05보다 크다.
- 유의수준 5%에서 귀무가설을 기각할 수 없다.
- **[결론] 후판 지시길이는 불량률은 관련이 없다.**

탐색적 분석 _(5) 변수별 검정

- Hsb와 scale의 독립성 가설검정 → HSB(Hot Scale Bracker)별 scale 차이 有

Hsb




```
chi-squared statistic: 106.21295741716654
p-value: 6.622223001546653e-25
degrees of freedom: 1
```

귀무가설	(H0): ‘hsb’ 변수와 ‘scale’ 변수는 독립이다.
대립가설	(H1): ‘hsb’ 변수와 ‘scale’ 변수는 독립적이지 않다.

[결론] : 대립가설 채택

•“hsb와 불량률은 관련이 있다.”

[근거]

•P-값이 0.05보다 작다.

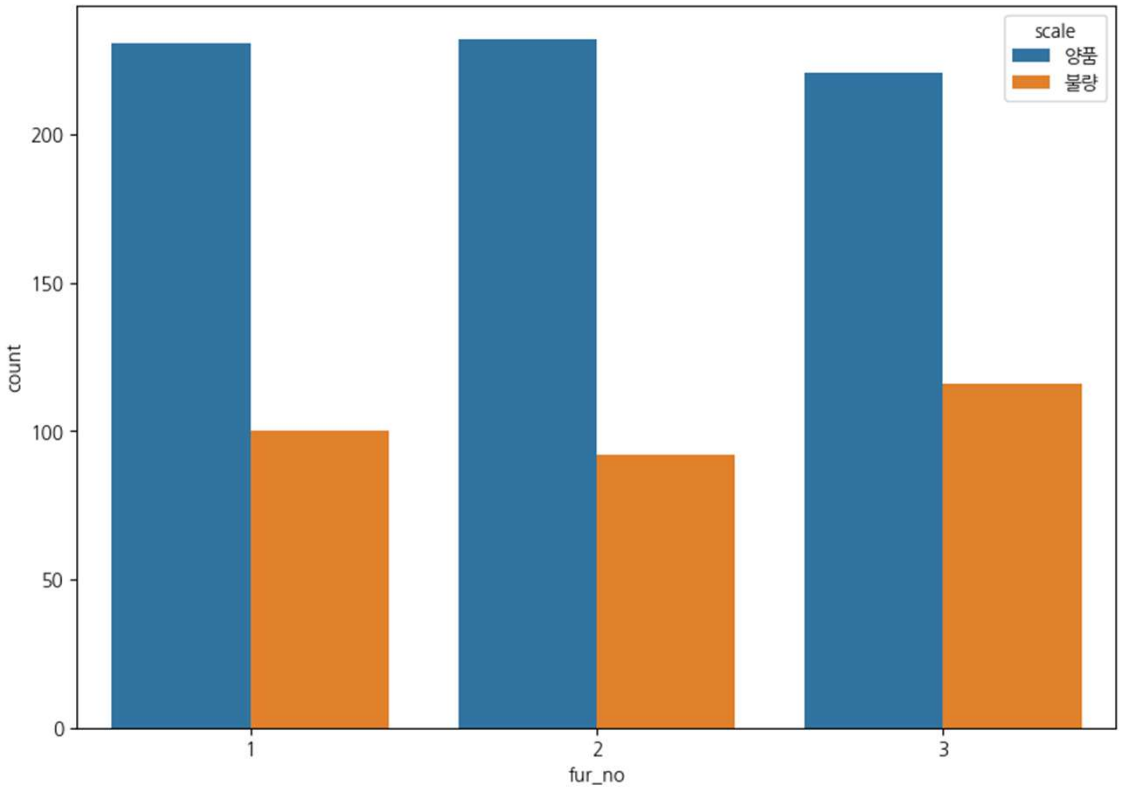
•유의수준 5%에서 귀무가설을 기각할 수 있다.

•그러므로, hsb와 불량률은 관련이 있다.

탐색적 분석 _(5) 변수별 검정

- Fur_no와 scale의 독립성 가설검정 → 가열로 호기별 scale 차이 無

Fur_no




```
chi-squared statistic: 2.964674399054447
p-value: 0.22710627493767527
degrees of freedom: 2
```

귀무가설	(H0): 'fur_no' 변수와 'scale' 변수는 독립이다.
대립가설	(H1): 'fur_no' 변수와 'scale' 변수는 독립적이지 않다.

[결론] : 귀무가설 채택

•"fur_no와 불량률은 관련이 없다."

[근거]

•P-값이 0.05보다 크다.

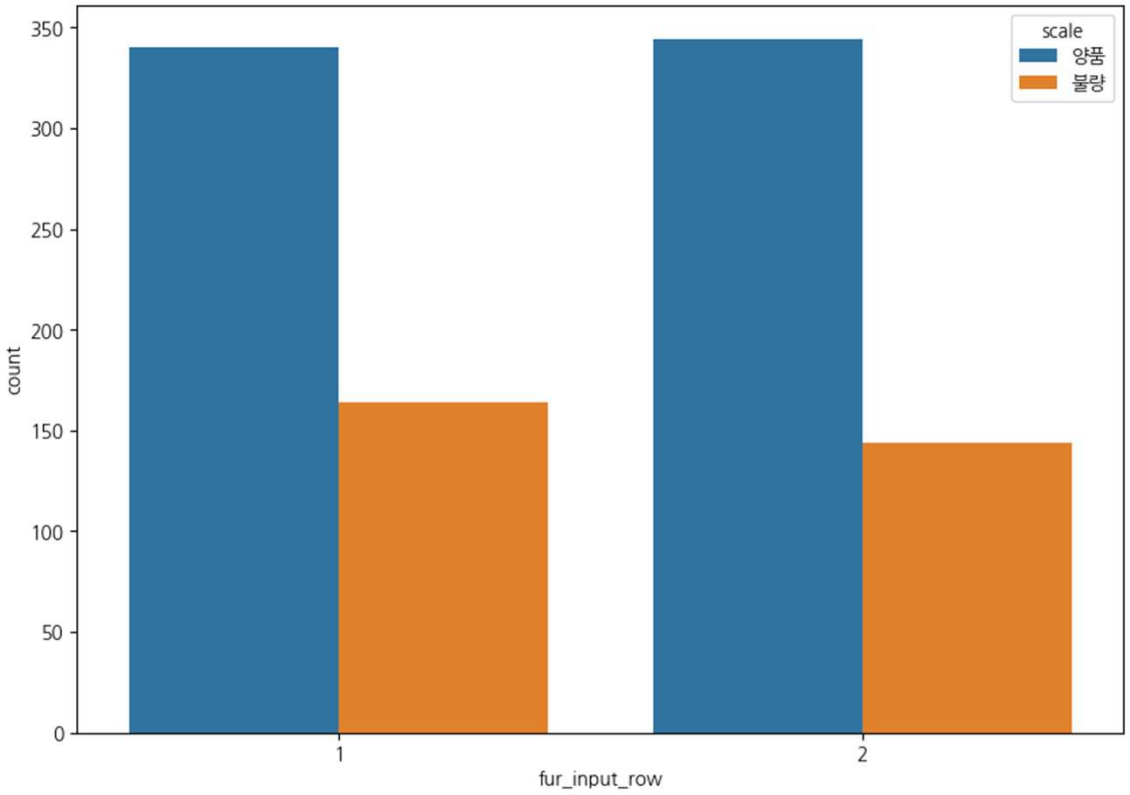
•유의수준 5%에서 귀무가설을 기각할 수 없다.

•그러므로, hsb와 불량률은 관련이 없다.

탐색적 분석 _(5) 변수별 검정

- Fur_input_row와 scale의 독립성 가설검정 → 가열로 장입열별 scale 차이 無

Fur_input_row




```
chi-squared statistic: 0.9274125630590576
p-value: 0.33553617272380376
degrees of freedom: 1
```

귀무가설	(H0): 'fur_input_row' 변수와 'scale' 변수는 독립이다.
대립가설	(H1): 'fur_input_row' 변수와 'scale' 변수는 독립적이지 않다.

[결론] : 귀무가설 채택

•"fur_input_row와 불량률은 관련이 없다."

[근거]

•P-값이 0.05보다 크다.

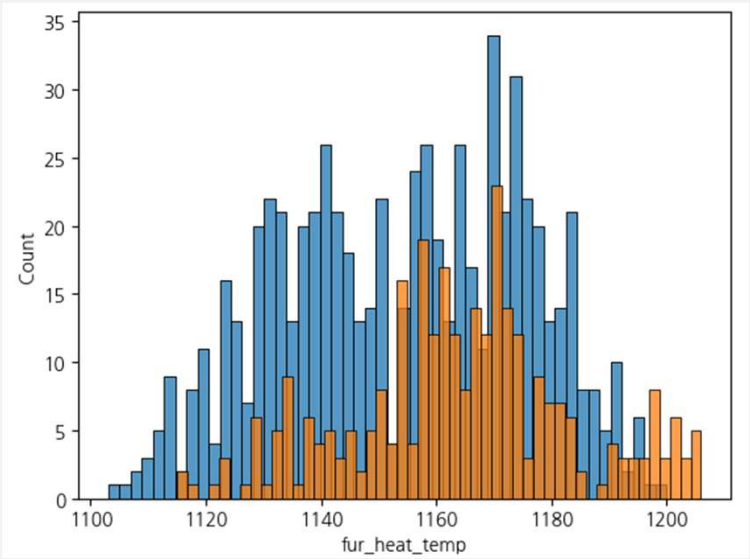
•유의수준 5%에서 귀무가설을 기각할 수 없다.

•그러므로, hsb와 불량률은 관련이 없다.

탐색적 분석 _(5) 변수별 검정

Fur_heat_temp의 독립성 가설검정 → 가열로 가열대 소재 온도별 scale 차이 有

Fur_heat_temp




```
Optimization terminated successfully.
         Current function value: 0.593556
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:                  scale   No. Observations:                  992
Model:                          Logit   Df Residuals:                      990
Method:                           MLE   Df Model:                            1
Date:                Tue, 28 May 2024   Pseudo R-squ.:                0.04186
Time:                        19:43:34   Log-Likelihood:                -588.81
converged:                       True   LL-Null:                       -614.53
Covariance Type:            nonrobust   LLR p-value:                7.353e-13
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         28.9005      4.082      7.081      0.000      20.901      36.900
fur_heat_temp -0.0242      0.004     -6.898      0.000      -0.031      -0.017
==============================================================================
```

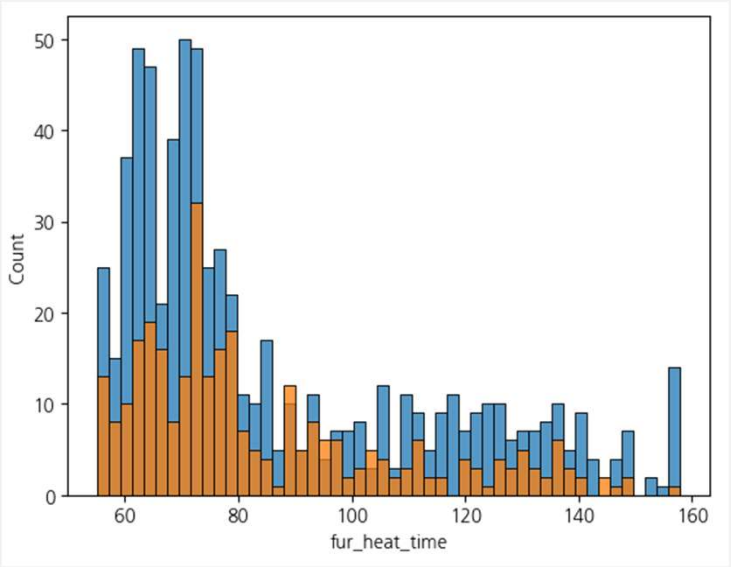
귀무가설	(H0): 'fur_heat_temp' 변수와 'scale' 변수는 독립이다.
대립가설	(H1): 'fur_heat_temp' 변수와 'scale' 변수는 독립적이지 않다.

- P-값이 0.05보다 작다.
- 유의수준 5%에서 귀무가설을 기각할 수 있다.
- **[결론] fur_heat_temp와 불량률은 관련이 있다.**

탐색적 분석 _(5) 변수별 검정

Fur_heat_time의 독립성 가설검정 → 가열로 가열대 재로 시간별 scale 차이 無

Fur_heat_time




```
Optimization terminated successfully.
         Current function value: 0.618841
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:                  scale   No. Observations:                  992
Model:                          Logit   Df Residuals:                      990
Method:                           MLE   Df Model:                            1
Date:                Tue, 28 May 2024   Pseudo R-squ.:                0.001043
Time:                        19:43:34   Log-Likelihood:                -613.89
converged:                       True   LL-Null:                       -614.53
Covariance Type:            nonrobust   LLR p-value:                    0.2576
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.5431      0.236      2.306      0.021       0.081       1.005
fur_heat_time  0.0030      0.003      1.125      0.261      -0.002       0.008
==============================================================================
```

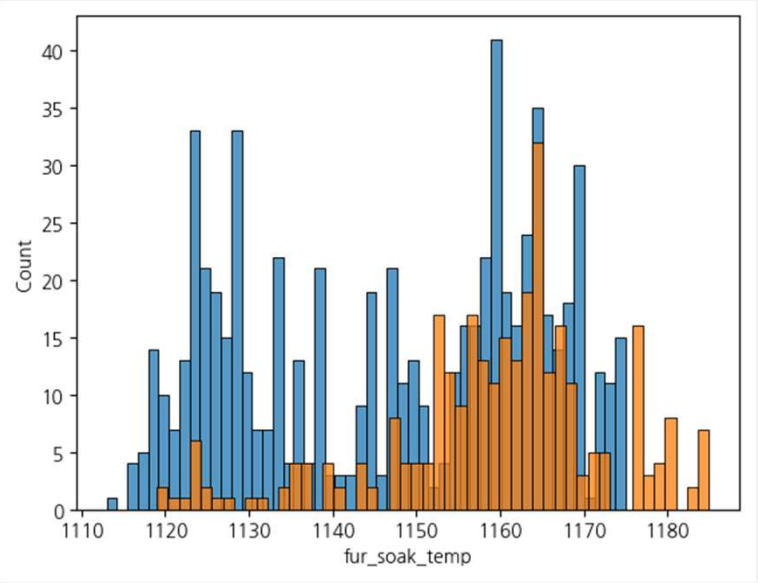
귀무가설	(H0): fur_heat_time’ 변수와 ‘scale’ 변수는 독립이다.
대립가설	(H1): ‘fur_heat_time’ 변수와 ‘scale’ 변수는 독립적이지 않다.

- P-값이 0.05보다 크다.
- 유의수준 5%에서 귀무가설을 기각할 수 없다.
- **[결론] fur_heat_time과 불량률은 관련이 없다.**

탐색적 분석 _(5) 변수별 검정

Fur_soak_temp의 독립성 가설검정 → 가열로 균열대 소재 온도별 scale 차이 有

Fur_soak_emp



Optimization terminated successfully.
Current function value: 0.560132
Iterations 6

Logit Regression Results

Dep. Variable:	scale	No. Observations:	992
Model:	Logit	Df Residuals:	990
Method:	MLE	Df Model:	1
Date:	Tue, 28 May 2024	Pseudo R-squ.:	0.09581
Time:	19:43:35	Log-Likelihood:	-555.65
converged:	True	LL-Null:	-614.53
Covariance Type:	nonrobust	LLR p-value:	1.956e-27

	coef	std err	z	P>|z|	[0.025	0.975]
const	57.1213	5.788	9.869	0.000	45.778	68.465
fur_soak_temp	-0.0488	0.005	-9.755	0.000	-0.059	-0.039

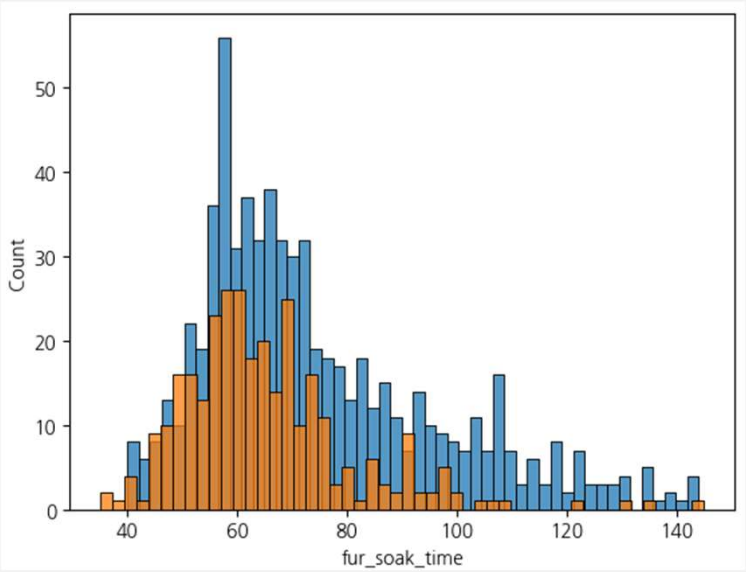
귀무가설	(H0): fur_soak _temp' 변수와 'scale' 변수는 독립이다.
대립가설	(H1): 'fur_soak_temp' 변수와 'scale' 변수는 독립적이지 않다.

- P-값이 0.05보다 작다.
- 유의수준 5%에서 귀무가설을 기각할 수 있다.
- **[결론] fur_soak_temp와 불량률은 관련이 있다.**

탐색적 분석 _(5) 변수별 검정

Fur_soak_time의 독립성 가설검정 → 가열로 균열대 재로 시간별 scale 차이 有

Fur_soak_time



Optimization terminated successfully.
Current function value: 0.592508
Iterations 6

Logit Regression Results

Dep. Variable:	scale	No. Observations:	992
Model:	Logit	Df Residuals:	990
Method:	MLE	Df Model:	1
Date:	Tue, 28 May 2024	Pseudo R-squ.:	0.04355
Time:	19:43:35	Log-Likelihood:	-587.77
converged:	True	LL-Null:	-614.53
Covariance Type:	nonrobust	LLR p-value:	2.552e-13

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1401	0.290	-3.933	0.000	-1.708	-0.572
fur_soak_time	0.0279	0.004	6.664	0.000	0.020	0.036

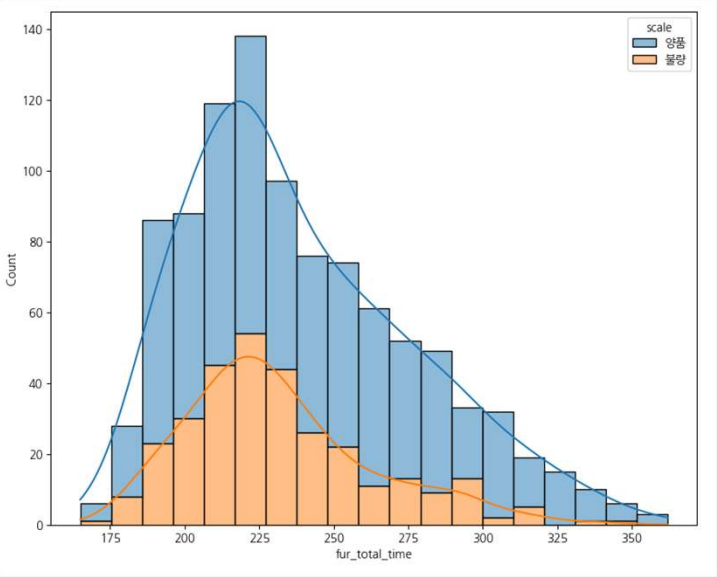
귀무가설	(H0): fur_soak_time' 변수와 'scale' 변수는 독립이다.
대립가설	(H1): 'fur_soak_time' 변수와 'scale' 변수는 독립적이지 않다.

- P-값이 0.05보다 작다.
- 유의수준 5%에서 귀무가설을 기각할 수 있다.
- **[결론] fur_soak_time과 불량률은 관련이 있다.**

탐색적 분석 _(5) 변수별 검정

Fur_total_time의 독립성 가설검정 → 가열로 총 재로 시간별 scale 차이 有

Fur_total_time




```
Optimization terminated successfully.
         Current function value: 0.611063
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:                  scale   No. Observations:                  992
Model:                          Logit   Df Residuals:                      990
Method:                           MLE   Df Model:                            1
Date:                Tue, 28 May 2024   Pseudo R-squ.:                0.01360
Time:                        20:53:41   Log-Likelihood:                -606.17
converged:                       True   LL-Null:                       -614.53
Covariance Type:            nonrobust   LLR p-value:                 4.347e-05
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.0046      0.453     -2.216      0.027      -1.893      -0.116
fur_total_time   0.0076      0.002      3.985      0.000       0.004       0.011
==============================================================================
```

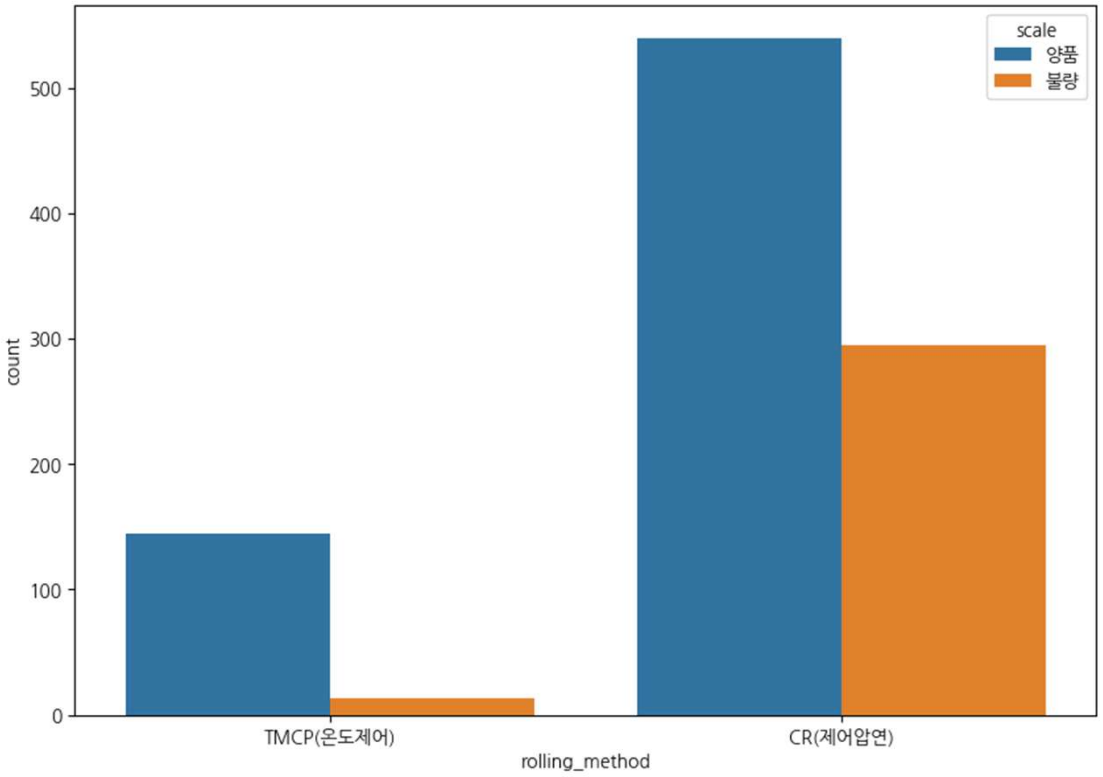
귀무가설	(H0): ‘fur_total_time’ 변수와 ‘scale’ 변수는 독립이다.
대립가설	(H1): ‘fur_total_time’ 변수와 ‘scale’ 변수는 독립적이지 않다.

- P-값이 0.05보다 작다.
- 유의수준 5%에서 귀무가설을 기각할 수 있다.
- **[결론] fur_total_time과 불량률은 관련이 있다.**

탐색적 분석 _(5) 변수별 검정

Rolling_method의 독립성 가설검정 → 압연방법별 scale 차이 有

Rolling_method




```
chi-squared statistic: 44.4571776284064
p-value: 2.5998205948808205e-11
degrees of freedom: 1
```

귀무가설	(H0): ‘rolling_method’ 변수와 ‘scale’ 변수는 독립이다.
대립가설	(H1): ‘rolling_method’ 변수와 ‘scale’ 변수는 독립적이지 않다.

[결론] : 대립가설 채택

•“rolling_method와 불량률은 관련이 있다.”

[근거]

•P-값이 0에 수렴한다.

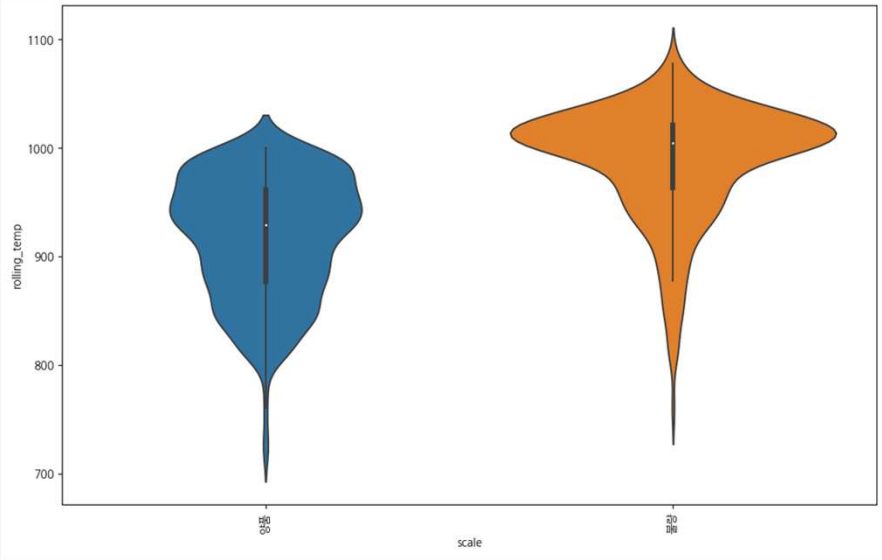
•유의수준 5%에서 귀무가설을 기각할 수 있다.

•그러므로, rolling_method와 불량률은 관련이 있다.

탐색적 분석 _(5) 변수별 검정

Rolling_temp의 독립성 가설검정 → 압연온도별 scale 차이 有

Rolling_temp




```
Optimization terminated successfully.
         Current function value: 0.455415
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                  scale   No. Observations:                  992
Model:                          Logit   Df Residuals:                      990
Method:                           MLE   Df Model:                            1
Date:                Tue, 28 May 2024   Pseudo R-squ.:                 0.2649
Time:                        19:43:35   Log-Likelihood:                -451.77
converged:                       True   LL-Null:                       -614.53
Covariance Type:            nonrobust   LLR p-value:                9.096e-73
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         27.1753      1.945     13.973      0.000      23.364      30.987
rolling_temp  -0.0275      0.002    -13.754      0.000      -0.031      -0.024
==============================================================================
```

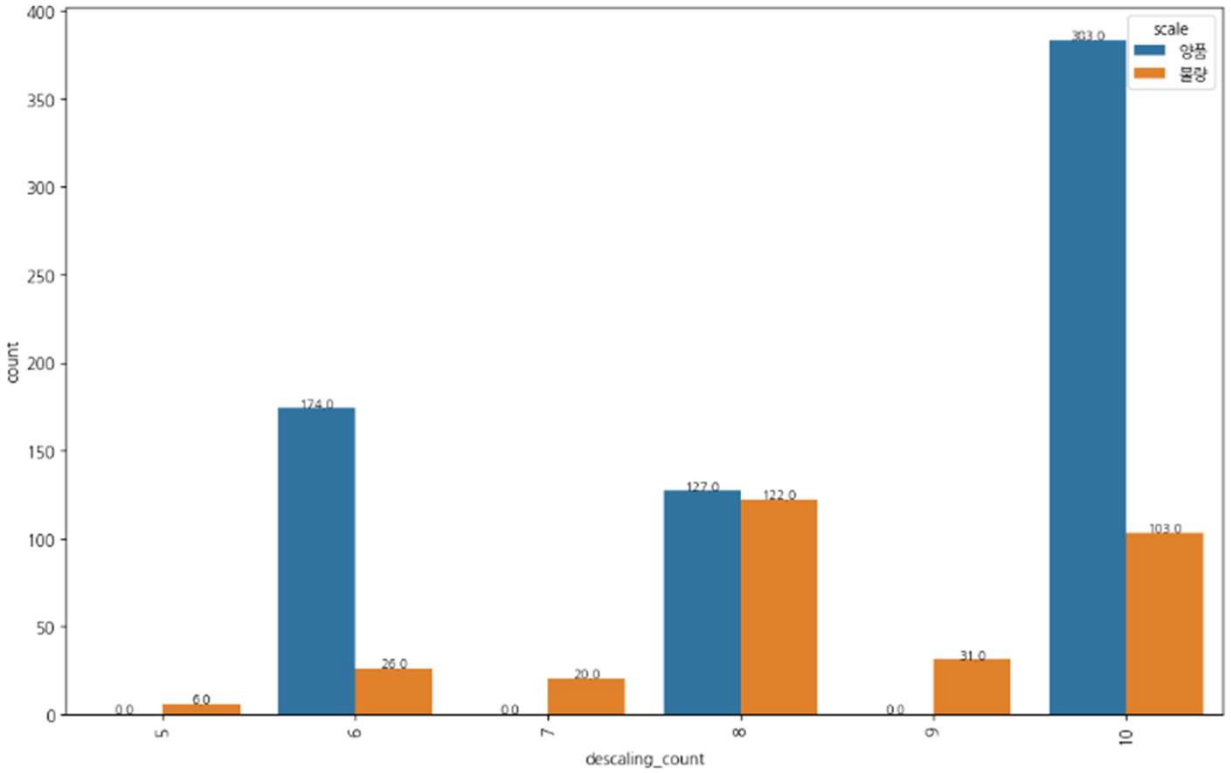
귀무가설	(H0): rolling_temp' 변수와 'scale' 변수는 독립이다.
대립가설	(H1): 'rolling_temp' 변수와 'scale' 변수는 독립적이지 않다.

- P-값이 0.05보다 작다.
- 유의수준 5%에서 귀무가설을 기각할 수 있다.
- **[결론] rolling_temp와 불량률은 관련이 있다.**

탐색적 분석 _(1) 그래프 탐색

- Descaling_count의 독립성 가설검정 → 압연 Descaling 횟수별 scale 차이 有

Descaling_count




```
chi-squared statistic: 216.52886196955083
p-value: 8.230963010047285e-45
degrees of freedom: 5
```

귀무가설	(H0): ‘descaling_count’ 변수와 ‘scale’ 변수는 독립이다.
대립가설	•(H1): ‘descaling_count’ 변수와 ‘scale’ 변수는 독립적이지 않다.

[결론] : 대립가설 채택

- “연료 타입 별과 불량률은 관련이 있다.”

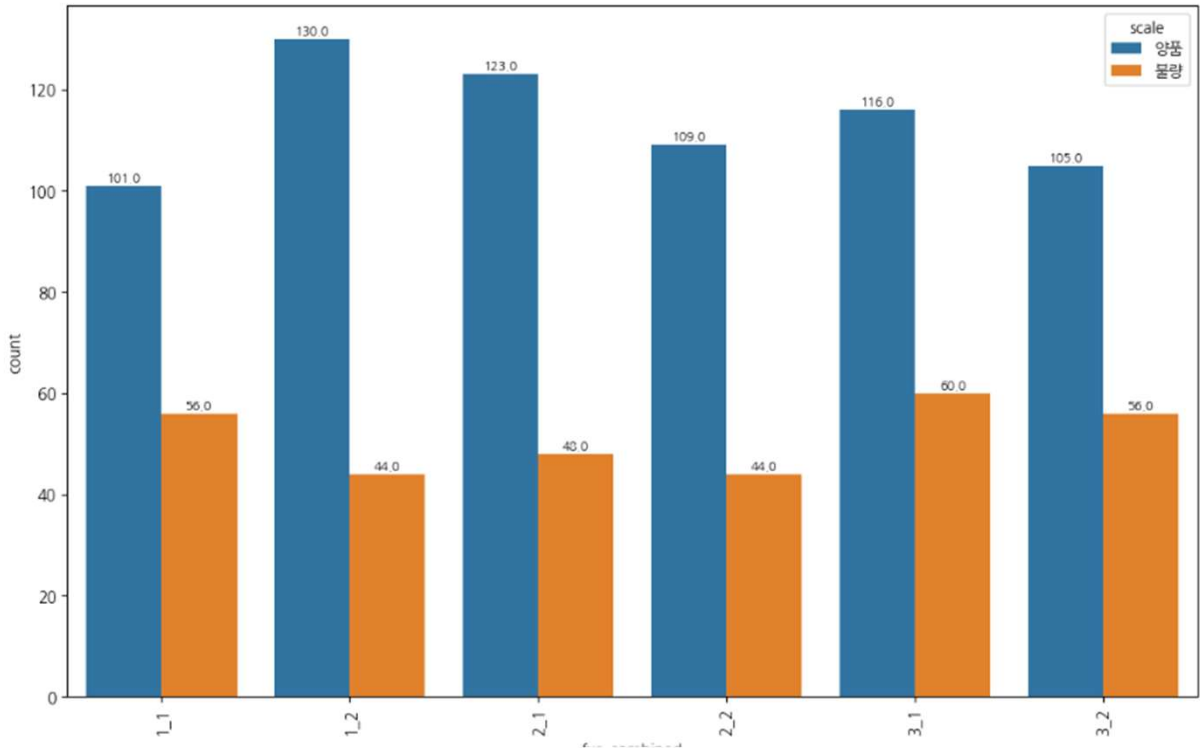
[근거]

- P-값이 0에 수렴한다.
- 유의수준 5%에서 귀무가설을 기각할 수 있다.
- 그러므로, descaling_count와 불량률은 관련이 있다.

탐색적 분석 _(1) 그래프 탐색

- Fur_combined의 scale차 가설검정 → 가열로 호기와 장입열별 scale 차이 有

Fur_combined




```
chi-squared statistic: 14.030960672980973
p-value: 0.002863311465440309
degrees of freedom: 3
```

귀무가설	(H0): 'scale'변수와 'fur_combined' 변수 독립적
대립가설	(H1): 'scale'변수와 'fur_combined' 변수 독립적이지 않음

[결론] : 대립가설 채택

- "가열로 호기와 장입열별 불량률은 관련이 있다."

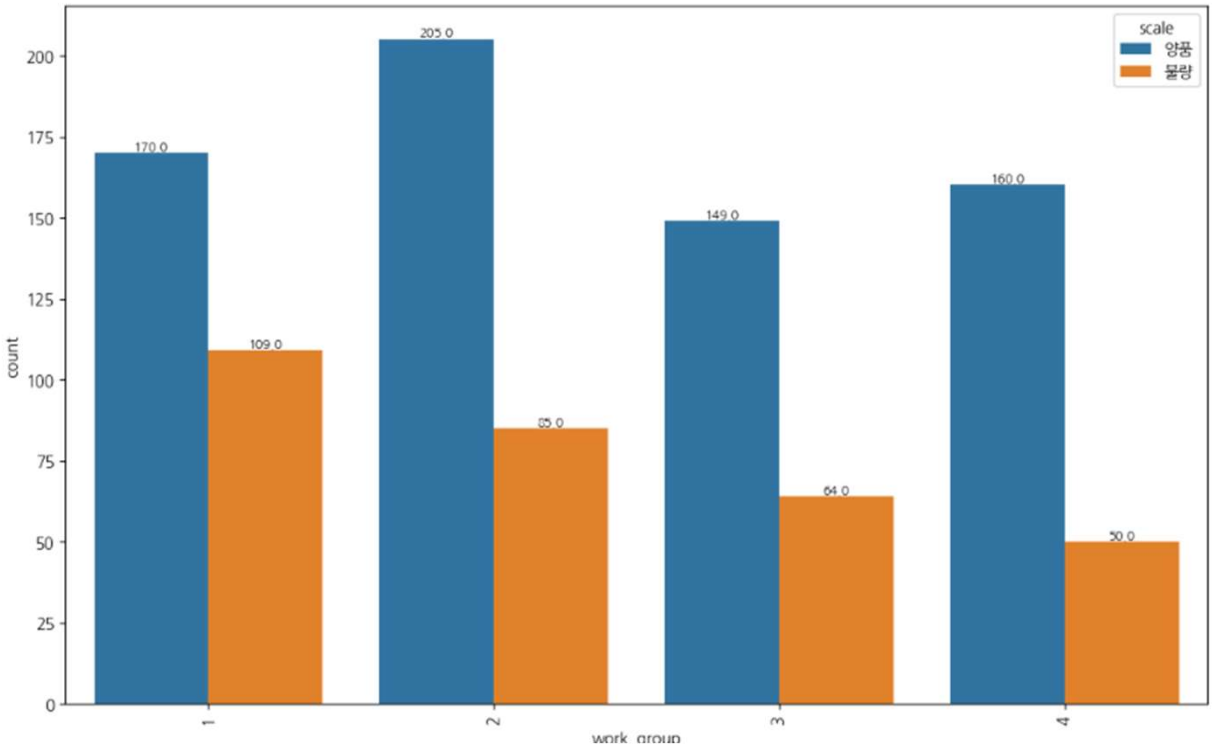
[근거]

- P-값이 0.2092로 0.05보다 크다.
- 유의수준 5%에서 귀무가설을 기각할 수 없다.
- 그러므로, 가열호 호기와 장입열별 불량률은 관련이 있다.

탐색적 분석 _(1) 그래프 탐색

- Work_group의 독립성 가설검정 → 가열로 호별 scale 차이 有

Work_group




```
chi-squared statistic: 14.030960672980973
p-value: 0.00286331146544030 9
degrees of freedom: 3
```

귀무가설	(H0): 'scale'변수와'work_group' 변수 독립적
대립가설	(H1): 'scale'변수와'work_group' 변수 독립적이지 않음

[결론] : 대립가설 채택

- "work_group과 불량률은 관련이 있다."

[근거]

- p-value가 0.05보다 작다.
- 유의수준 5%에서 귀무가설을 기각할 수 있다.
- 그러므로, work_group과 불량률은 관련이 있다.

- 기본 모델을 활용하여 score 비교

1. Decision Tree Classifier (의사결정나무)

```
from sklearn.tree import DecisionTreeClassifier #의사결정나무

#기본 옵션으로 모델 생성
tree_uncustomized = DecisionTreeClassifier(random_state = 1234)
tree_uncustomized.fit(train_x, train_y)

#train 모델 정확도
print('Accuracy on training set : {:.3f}'.format(tree_uncustomized.score(train_x, train_y)))

#test 모델 정확도
print('Accuracy on test set : {:.3f}'.format(tree_uncustomized.score(test_x, test_y)))
#의사결정나무 정확도 산출
```

```
Accuracy on training set : 1.000
Accuracy on test set : 0.990
```

>>> Test Score: 0.990

2. Gradient Boosting Classifier (그래디언트 부스팅)

```
from sklearn.ensemble import GradientBoostingClassifier #그래디언트 부스팅

tree_uncustom = GradientBoostingClassifier(random_state=1234)
tree_uncustom.fit(train_x, train_y)

print("train data의 score: {:.3f}".format(tree_uncustom.score(train_x, train_y)))
print("test data의 score: {:.3f}".format(tree_uncustom.score(test_x, test_y)))
#그래디언스 부스팅 정확도 산출
```

```
train data의 score: 1.000
test data의 score: 0.990
```

>>> Test Score: 0.990

3. Random Forest Classifier (랜덤포레스트)

```
from sklearn.ensemble import RandomForestClassifier #랜덤포레스트

rf_uncustomized = RandomForestClassifier(random_state = 1234)
rf_uncustomized.fit(train_x,train_y)
print("Score on training set: {:.3f}".format(rf_uncustomized.score(train_x,train_y)))
print("Score on test set: {:.3f}".format(rf_uncustomized.score(test_x,test_y)))
#랜덤포레스트 정확도 산출
```

```
Score on training set: 1.000
Score on test set: 0.960
```

>>> Test Score: 0.960

4. XG Booster

```
from xgboost import XGBClassifier #XGBoost


xgboost_uncust = XGBClassifier(random_state=1234)
xgboost_uncust.fit(train_x, train_y)

print("train data의 score: {:.3f}".format(xgboost_uncust.score(train_x, train_y)))
print("test data의 score: {:.3f}".format(xgboost_uncust.score(test_x, test_y)))
```

```
train data의 score: 1.000
test data의 score: 0.993
```

>>> Test Score: 0.993

-	Decision Tree	Random Forest	Gradient Boost	XG Boost
Train Score	100.0	100.0	100.0	100.0
Test Score	99.0	96.0	99.0	99.3

→ 그래디언트 부스트 선정

최종 모델링_XG Boost

- 최종 모델으로 XG Boost 선정 (Score = 0.993)

[최종 모델링 결과]

조기 중단 모델 평가 : Test

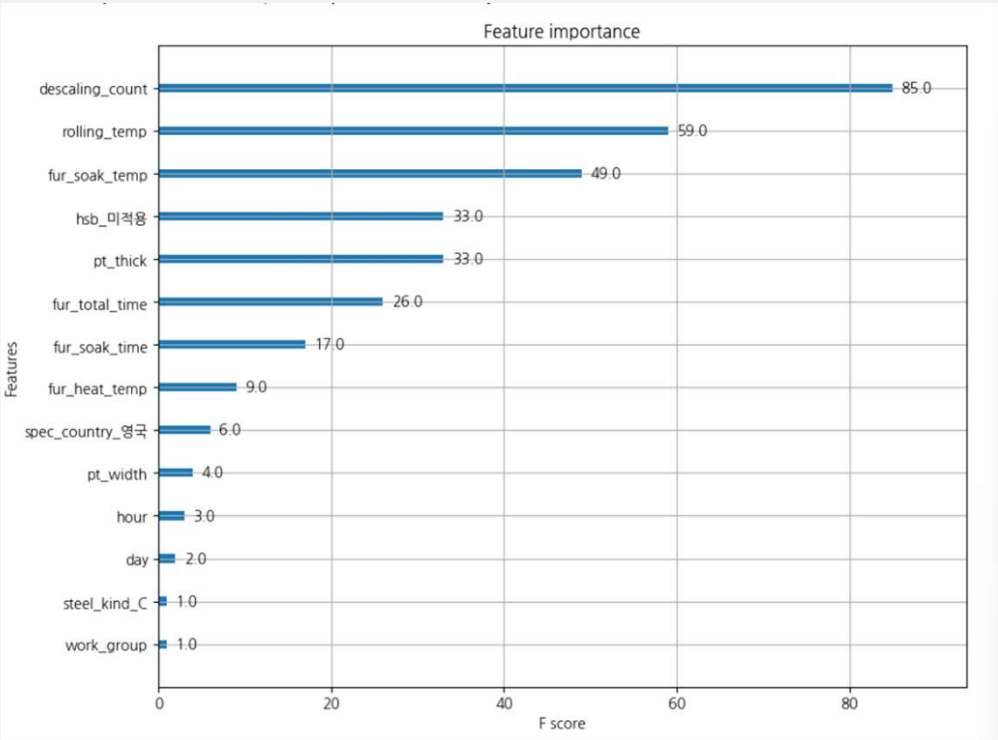
오차행렬:
[[90 4]
[0 204]]

정확도: 0.9866
정밀도: 0.9808
재현율: 1.0000
F1 : 0.9903
AUC : 0.9787

	precision	recall	f1-score	support
0	1.00	0.96	0.98	94
1	0.98	1.00	0.99	204
accuracy			0.99	298
macro avg	0.99	0.98	0.98	298
weighted avg	0.99	0.99	0.99	298

- 정확도(98.66%) → 전체 데이어의 98.66% 정확히 예측
- 정밀도 (0: 100%, 1: 98%) → 1로 예측한 것 중 98%가 실제로 1
- 재현율 (0: 96%, 1: 100%) → 0으로 예측한 것 중 96%가 실제로 0
- F1점수 (0: 98%, 1: 99%) → 0과 1에 두 값에 대한 예측이 얼마나 균형있게 높은지
- AUC (97.87%) → 모델의 성능이 매우 높은것으로 판단

[변수 중요도]



- 1순위: descaling count (85.0%) → 모델에서 가장 중요한 변수 (유지보수, 적정온도 설정 필요)
- 2순위: rolling temp (59.0%) → 적정 롤링 온도와 스케일 불량발생 상관성
- 3순위: hsb_미적용 (33.0%) → hsb 적용 여부에 따라 불량 발생
- 4순위: pt_thick (33.0%) → 후판이 두꺼워질수록 스케일 불량 가능성 감소

결론

- 핵심영향요인 도출 결과 후판 제품 scale 불량은 *HSB, 가열로 담금, 디스케일링, 압연 공정*에서 온도 이상(고온)때문임을 식별함

열연 스케일 제거기 (HSB)	가열로 담금
HSB 적용을 통해 재료를 균일하게 처리함으로써 제품의 일관성 보장 필요	적정 온도 유지 및 관리를 통해 표면 품질을 강화와 탈탄 현상 방지 재료 가열 온도를 적정수준으로 유지하여 결함 발생 최소화
디스케일링	**압연공정**
정기적 유지관리를 통한 장비 노후화 방지 적절한 온도 설정으로 스케일 형성 억제	롤링 온도 과열로 인한 변형 예방 적정 온도 설정 및 모니터링을 통한 표면 결함 최소화

① 적절한 온도를 설정, *② 충분한 장비 유지보수 탈탄 및 표면 품질 저하 방지*, *③ 압연 공정 개선*을 통해 Scale 불량률을 0%로 예방하고 품질 향상을 도모할 수 있을 것으로 기대

개선안 및 느낀점

With POSCO

도메인 지식

- 조원들과 자료 조사와 분석 계획을 수립하는 과정에서 도메인 지식과 비교하며 데이터를 분석하였다. 그 중에서도 Descaling count 데이터를 분석할 때, 홀수 횟수에서 100%의 불량률이 나온다는 점이 도메인 지식과는 상당히 위배되는 결과여서 무척 당황하였다.

데이터 전처리

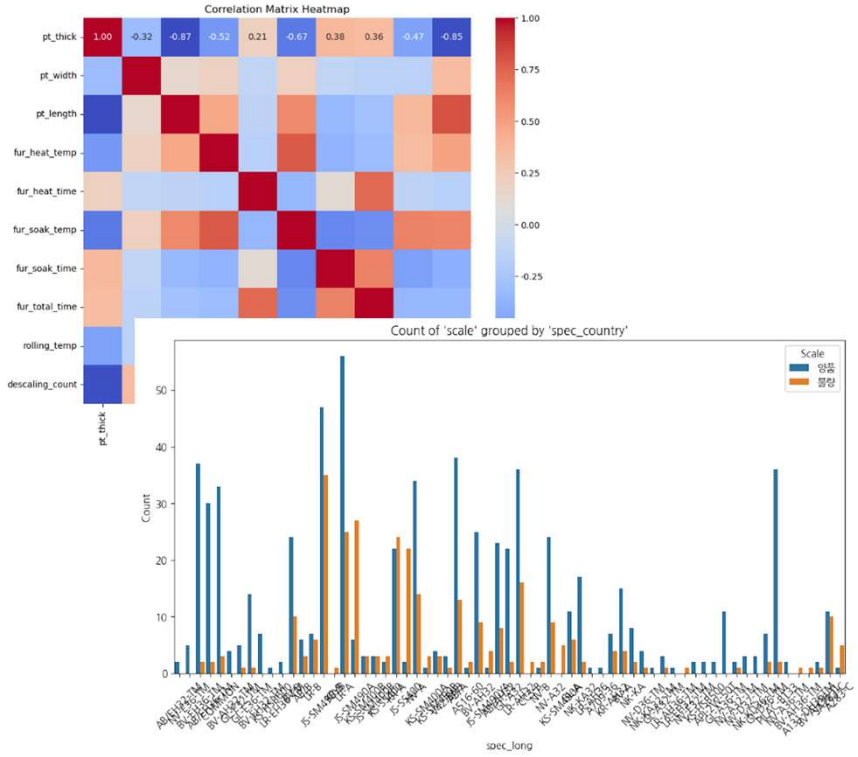
데이터 전처리 과정에서 발견한 이상치를 제거할 것인지 대체할 것 인지에 대해 조원들과 오랜 시간 고민하였다. 각 데이터의 평균과 표준편차를 자세히 비교하면서 조원들과의 토의를 통해 합리적인 의사 결정을 할 수 있었다.

모델링 및 파라미터 선정

- 모델링 과정에서 하이퍼 파라미터를 선정하며 모델을 train 했지만 예상보다 score 값이 좋지 않아서 모델 성능 향상을 위해 여러 번 파라미터를 조절하고 값을 변경하였고 결과 score를 비교하여 모델의 성능을 향상시키기 위해 노력하였다.
- 모델링 성능 분석 과정에서 Grid Search CV와 하이퍼 파라미터를 선정한 트리 모델들과 비교하였을 때 Grid Search CV 모델의 성능보다 직접 하이퍼 파라미터를 선정한 모델의 성능이 더 좋아서 오히려 모델의 성능을 의심하였다.

그래프 탐색

데이터의 분포와 상관성을 표현하는 그래프를 그리는 과정에서 가장 가독성이 높고 해석력이 좋은 그래프를 만들고 싶어 여러 가지 방법을 시도해보았다. 모델에 따른 다중공선성 검토와 같이 필요 과정에 대한 이해가 초기에 미흡하여 시행착오를 겪었다. 그 결과 아래와 같은 그래프를 작성하였으나 최종 코드에는 제외하였다.



핵심인자 정리 템플릿(참조용)_1/2

With POSCO

변수	변수설명	변수역할	변수형태	분석 제외사유
plate_no	Plate번호	ID	범주형	라벨링 해주기 위해 생성된 값으로 내포된 의미가 없기 때문에
rolling_date	열연작업시각	날짜	연속형	열연작업시각에 포함되어있는 일, 시각을 구분하여 분석에 적용.
scale	Scale(산화철) 불량	목표변수	범주형	
spec_long	제품 규격	설명변수	범주형	제품 규격 첫 두글자가 기준국과 관련이 되어있는것을 확인, 설명변수간의 상관성을 고려해서 삭제예정
spec_country	제품 규격 기준국	설명변수	범주형	
steel_kind	강종	설명변수	범주형	
pt_thick	Plate(후판) 지시두께(mm)	설명변수	연속형	
pt_width	Plate(후판) 지시폭(mm)	설명변수	연속형	
pt_length	Plate(후판) 지시길이(mm)	설명변수	연속형	P-value 값이 0.05보다 커서 변수가 유의미하지 않은 것으로 판단.
hsb	HSB(Hot Scale Braker)적용여부	설명변수	범주형	
fur_no	가열로 호기	설명변수	범주형	P-value 값이 0.05보다 커서 변수가 유의미하지 않은 것으로 판단.
fur_input_row	가열로 장입열	설명변수	범주형	P-value 값이 0.05보다 커서 변수가 유의미하지 않은 것으로 판단.
fur_heat_temp	가열로 가열대 소재온도(℃)	설명변수	연속형	
fur_heat_time	가열로 가열대 재로시간(분)	설명변수	연속형	
fur_soak_temp	가열로 균열대 소재온도(℃)	설명변수	연속형	
fur_soak_time	가열로 균열대 재로시간(분)	설명변수	연속형	
fur_total_time	가열로 총 재로시간(분)	설명변수	연속형	
rolling_method	압연방법	설명변수	범주형	
rolling_temp	압연온도(℃)	설명변수	연속형	
descaling_count	압연Descaling 횟수	설명변수	연속형	
work_group	작업조	설명변수	범주형	
rolling_day	rolling_date의 일	파생변수	범주형	
rolling_hour	rolling_date의 시각	파생변수	범주형	

핵심인자 정리 템플릿(참조용)_2/2

변수	탐색적 기법			모델링 기법				총점	선정 (사유)
	그래프	검정	상관분석	회귀분석	DT	RF	GB		
plate_no	x	x	x	x	x	x	x	0	
rolling_date	x	x	x	x	x	x	x	0	
scale	o	o	o	o	o	o	o	7	
spec_long	o	x	x	x	x	x	x	1	
spec_country	o	o	o	o	o	o	o	7	가설검정을 통해 scale값의 변화에 유의미한 것이 확인 되어서
steel_kind	o	o	o	o	o	o	o	7	가설검정을 통해 scale값의 변화에 유의미한 것이 확인 되어서
pt_thick	o	o	o	o	o	o	o	7	가설검정을 통해 scale값의 변화에 유의미한 것이 확인 되어서
pt_width	o	o	o	o	o	o	o	7	가설검정을 통해 scale값의 변화에 유의미한 것이 확인 되어서
pt_length	o	o	x	x	x	x	x	2	
hsb	o	o	o	o	o	o	o	7	가설검정을 통해 scale값의 변화에 유의미한 것이 확인 되어서
fur_no	o	o	x	x	x	x	x	2	
fur_input_row	o	o	x	x	x	x	x	2	
fur_heat_temp	o	o	o	o	o	o	o	7	가설검정을 통해 scale값의 변화에 유의미한 것이 확인 되어서
fur_heat_time	o	o	o	o	o	o	o	7	가설검정을 통해 scale값의 변화에 유의미한 것이 확인 되어서
fur_soak_temp	o	o	o	o	o	o	o	7	가설검정을 통해 scale값의 변화에 유의미한 것이 확인 되어서
fur_soak_time	o	o	o	o	o	o	o	7	가설검정을 통해 scale값의 변화에 유의미한 것이 확인 되어서
fur_total_time	o	o	o	o	o	o	o	7	가설검정을 통해 scale값의 변화에 유의미한 것이 확인 되어서
rolling_method	o	o	o	o	o	o	o	7	가설검정을 통해 scale값의 변화에 유의미한 것이 확인 되어서
rolling_temp	o	o	o	o	o	o	o	7	가설검정을 통해 scale값의 변화에 유의미한 것이 확인 되어서
descaling_count	o	o	o	o	o	o	o	7	가설검정을 통해 scale값의 변화에 유의미한 것이 확인 되어서
work_group	o	o	o	o	o	o	o	7	가설검정을 통해 scale값의 변화에 유의미한 것이 확인 되어서
rolling_day	o	o	o	o	o	o	o	7	가설검정을 통해 scale값의 변화에 유의미한 것이 확인 되어서
rolling_hour	o	o	o	o	o	o	o	7	가설검정을 통해 scale값의 변화에 유의미한 것이 확인 되어서

이름	역할
공통	데이터 분석 프로세스 설계, 모델링 성능 비교
손정우(조장)	분석 프로세스 검토, 변수 파생·변형, XG Boost 모델링 및 성능평가, 최종모델링
조현서	자료조사, 독립성 가설검정, 그래프 탐색, PPT
노성훈	자료조사, 데이터 전처리, 랜덤포레스트 모델링 및 성능평가
길배섭	데이터 전처리, 변수 파생·변형, 그래디언트 부스트 모델링 및 성능평가
정재원	자료조사, 데이터 전처리, 변수 파생·변형, 최종모델링, PPT

감사합니다

With POSCO

조현서, 노성훈, 길배섭, 정재원, 손정우