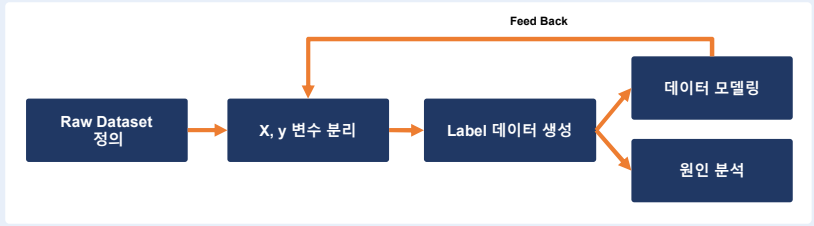
1. 데이터 개요

- **데이터 분석 Flow Chart**

1. Raw Dataset 정의
2. Target 설정
3. Target 모델링을 통한 원인 분석 변수 Target 선정
4. 해당 분석 Target 공정별로 반복



1. 데이터 개요

- **데이터 현황**

20240627_A3EA1_DATA_RG3-L.xlsx [**S360**]

S360 모터 Assly Line

20519행 & 61열		2024.01.07 ~ 2024.06.20	
X Input Features			Y Target Features
COGGING_HARMONIC ORDER 8의 인자		COGGING_HARMONIC ORDER 12 or 24의 인자	COGGING_TORQUE_mNm의 관련 결과
S32_Press_distance_X1	S34_Press distance X1	S230_QD_MAGN_CURRENT_KA	S360_QD_CW_COGGING_HARMONIC ORDER 8_mNm
S32_Press_distance_X2	S34_Press distance X2	S230_QD_MAGN_VOLTAGE_V	
S32_Press_distance_X3	S34_Press distance X3	S230_QD_TEMPERATURE	S360_QD_CW_COGGING_HARMONIC ORDER 12_mNm
S32_Press_force_F1	S34_Press force F1	S240_QD_STACK_A_AREASUM	
S32_Press_force_F2	S34_Press force F2	S240_QD_STACK_B_AREASUM	S360_QD_CW_COGGING_HARMONIC ORDER 24_mNm
S32_Press_force_F3	S34_Press force F3	S240_QD_STACK_A-B Skew	S360_QD_CCW_COGGING_HARMONIC ORDER 8_mNm
S70_Actual_Distance_X1_(mm)	S70_Pressing_Force_F1_(N)	S240_QD_STACK_A-C Skew	
S70_Actual_Distance_X2_(mm)	S70_Pressing_Force_F2_(N)	S240_QD_STACK_A Angle	S360_QD_CCW_COGGING_HARMONIC ORDER 12_mNm
S70_Actual_Distance_X3_(mm)	S70_Pressing_Force_F3_(N)	S240_QD_STACK_B Angle	
S70_Actual_Distance_X4_(mm)	S70_Pressing_Force_F4_(N)	S240_QD_STACK_C Angle	S360_QD_CCW_COGGING_HARMONIC ORDER 24_mNm
S70_Actual_Distance_X5_(mm)	S70_Pressing_Force_F5_(N)	S240_QD_STACK_B-C Skew	
S70_Actual_Distance_X6_(mm)	S70_Pressing_Force_F6_(N)		
S70_Heating_Temperature			
S71_Actual_Distance_X1_(mm)	S71_Pressing_Force_F1_(N)		
S71_Actual_Distance_X2_(mm)	S71_Pressing_Force_F2_(N)		
S71_Actual_Distance_X3_(mm)	S71_Pressing_Force_F3_(N)		
S71_Actual_Distance_X4_(mm)	S71_Pressing_Force_F4_(N)		
S71_Actual_Distance_X5_(mm)	S71_Pressing_Force_F5_(N)		
S71_Actual_Distance_X6_(mm)	S71_Pressing_Force_F6_(N)		
S71_Heating_Temperature			

2. 데이터 탐색 및 전처리

- **X, y 변수 분리 개념**

- 각 조립 공정별 결과치를 X변수로 정의
- 최종 공정의 결과치를 Y값으로 정의
- 해당 분류 과정을 토대로 모델링 진행
- **Y값의 값들을 예측하는데 어떤 X 인자가 영향을 주는지 파악하기 위함(원인분석)**

→ 예시

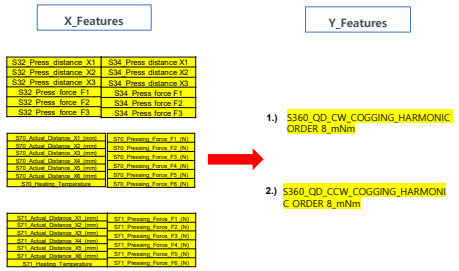
X_Features	Y_Features
'S240_QD_STACK_A_AREASUM', 'S240_QD_STACK_A-B Skew', 'S240_QD_STACK_A Angle', 'S240_QD_STACK_C Angle', 'S240_QD_STACK_B_AREASUM', 'S240_QD_STACK_A-C Skew', 'S240_QD_STACK_B Angle', 'S240_QD_STACK_B-C Skew	'S360_QD_CW_COGGING_HARMONIC ORDER 24_mNm

2. 데이터 탐색 및 전처리

- **병렬 조립공정**

- S32 – S70 – 공정검사 결과 [CW & CCW]
- S32 – S70 – 공정검사 결과 [CW & CCW]
- S32 – S71 – 공정검사 결과 [CW & CCW]
- S32 – S71 – 공정검사 결과 [CW & CCW]
- S34 – S70 – 공정검사 결과 [CW & CCW]
- S34 – S70 – 공정검사 결과 [CW & CCW]
- S34 – S71 – 공정검사 결과 [CW & CCW]
- S34 – S71 – 공정검사 결과 [CW & CCW]

→ **8개의 조합으로 데이터 분할**



8개의 조합 X 2 Target = 총 16개의 데이터 조합 생성

2. 데이터 탐색 및 전처리

- 기타 조립공정

X_Features

S230_QD_MAGN_CURRENT_KA
S230_QD_MAGN_VOLTAGE_V
S230_QD_TEMPERATURE
S240_QD_STACK_A_AREASUM
S240_QD_STACK_B_AREASUM
S240_QD_STACK_A-B Skew
S240_QD_STACK_A-C Skew
S240_QD_STACK_A Angle
S240_QD_STACK_B Angle
S240_QD_STACK_C Angle
S240_QD_STACK_B-C Skew
S360_QD_CW_COGGING_TORQUE_mNm(삭제)

Y_Features

1.) S360_QD_CW_COGGING_HARMONIC ORDER 12_mNm

2.) S360_QD_CW_COGGING_HARMONIC ORDER 24_mNm

3.) S360_QD_CCW_COGGING_HARMONIC ORDER 12_mNm

4.) S360_QD_CCW_COGGING_HARMONIC ORDER 24_mNm

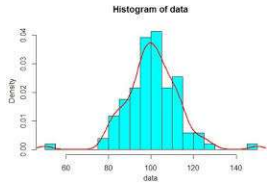
1개의 조합 X 4 Target = 데이터 4분할

2. 데이터 탐색 및 전처리

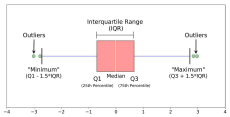
- Target Labeling

1. COGGING_TORQUE_mNm의 관련 결과 변수 대상

2. 각 변수들의 분포에 대한 임계값 선정 → [이상값 경계선]



Or



분포 기반 임계값 직접 탐색

주어진 임계값 활용 [13.75]

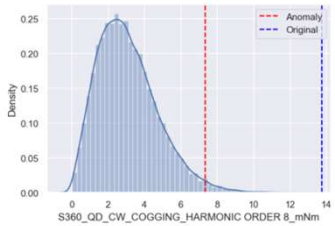
2. 데이터 탐색 및 전처리

- Target Labeling

3. 이상값 탐색 결과 예시

```
기존 임계값 ==> 13.75
0    20443
Name: S360_QD_CW_COGGING_HARMONIC ORDER 8_mNm, dtype: int64

Q3 + 1.5*IQR ==> 7.350775
0    20127
1      316
Name: S360_QD_CW_COGGING_HARMONIC ORDER 8_mNm, dtype: int64
불량률= 1.5457613853152673 %
```



기존의 임계값인 13.75를 기준으로 정상(0) & 불량(1) Labeling 수행할 경우 불량률은 거의 0%에 근접

임계값 수동 계산

사분위수 기반 이상치 탐지 & 이와 근사한 값을 설정한 이후 반복 예측 실험 수행

[어느 임계값 기준으로 분류 성능이 뛰어난지 실험 진행]

라벨링 표시

정상 → 0
불량 → 1

2. 데이터 탐색 및 전처리

- 데이터 정리

1. 병렬 조립공정

데이터 조합	결측 제거 개수 (행 기준)	최종 데이터 개수(행,열)	불량 품질 임계값	정상 Label 개수	불량 Label 개수	불량률(%)
S32-S70-CW-H8	15528	(4991, 20)	[13.75 → 7로 변경]	4892	99	≈ 1.98%
S32-S70-CCW-H8		(4991, 20)		4884	107	≈ 2.14%
S32-S71-CW-H8	15783	(4736, 20)		4671	65	≈ 1.37%
S32-S71-CCW-H8		(4736, 20)		4671	65	≈ 1.37%
S34-S70-CW-H8	15093	(5426, 20)		5364	62	≈ 1.14%
S34-S70-CCW-H8		(5426, 20)		5369	57	≈ 1.05%
S34-S71-CW-H8	15201	(5318, 20)		5286	32	≈ 0.60%
S34-S71-CCW-H8		(5318, 20)		5285	33	≈ 0.62%

Ex.) S360_QD_CW_COGGING_HARMONIC ORDER 8_mNm → CW-H8 명칭 약식

2. 데이터 탐색 및 전처리

- **데이터 정리**

2. 기타 조립공정

데이터 조합	결측 제거 개수 (행 기준)	최종 데이터 개수 (행,열)	불량 품질 임계값	정상 Label 개수	불량 Label 개수	불량률(%)
CW-H24	76	(20443, 12)	[13.75 → 11.59로 변경]	20292	151	≈ 0.73%
CCW-H24	76	(20443, 12)	[13.75 → 11.37로 변경]	20293	150	≈ 0.73%

** HARMONIC ORDER 12_mNm 제거 → 분석 수행 HARMONIC_ORDER 12에 대한 모델링 성능은 매우 저조

→ **최대 정확도 56% & 정상/불량 제대로 구별하지 못하는 수치 결과 [현재 개선가능성 X]**

2. 데이터 탐색 및 전처리

- **상관계수 탐색**

Ex.)

	S360_QD_CCW_COGGING_HARMONIC ORDER 24_mNm
S360_QD_CCW_COGGING_HARMONIC ORDER 24_mNm	1.000000
S230_QD_MAGN_CURRENT_KA	-0.314778
S230_QD_MAGN_VOLTAGE_V	-0.230545
S230_QD_TEMPERATURE	0.173634
S240_QD_STACK_A_AREASUM	-0.155477
S240_QD_STACK_B_AREASUM	0.130490
S240_QD_STACK_A-B Skew	0.125563
S240_QD_STACK_A-C Skew	0.005652
S240_QD_STACK_A Angle	0.077249
S240_QD_STACK_B Angle	0.065095
S240_QD_STACK_C Angle	0.045487
S240_QD_STACK_B-C Skew	-0.176264

1.) **대부분 공정 인자 변수들 & Target 변수들은 높은 상관관계를 보이지 않음**

2.) **Target Label의 불균형 비율 [정상 대비 매우 낮은 불량 개수] 아직 통제하지 못한 추가적인 변수들의 영향력 수집 데이터의 개수 등의 이유로 추정**

3. 모델링 과정 및 결과

- **데이터 학습 원리**

1. 분석 흐름도

데이터셋 개념 확립 → **데이터 전처리** → **예측 모델링 수립**

데이터 전처리
1.) X,y 변수 분할
2.) 병렬 / 기타 조립공정별 데이터셋 분할
3.) Target Labeling 진행 (정상 or 불량)

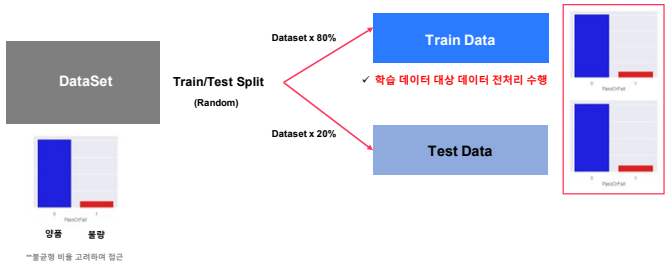
예측 모델링 수립
1.) 학습 / 검증 데이터 분할
2.) 학습 데이터 전처리 (불량비율 조정)
[**불량데이터 증강 Or 정상데이터 부분제거**]
3.) 변수 중요도 출력

3. 모델링 과정 및 결과

- **데이터 학습 원리**

2. ML(기계학습) 모델링의 통상적 개념

- **학습 데이터 [전체 데이터의 약 80%] & 검증 데이터 [전체 데이터의 약 20%]**
- **예측 성능 검증**



3. 모델링 과정 및 결과

데이터 전처리

1. Train Data 대상

- **Target Label의 불균형 비율 조정**
- **불량 데이터의 개수를 증강 (OverSampling) 또는 정상데이터의 개수를 제거 (UnderSampling)**
- **증강 & 삭감의 비율 임의 조정 가능**

기본적인 Random UnderSampling 진행

선택

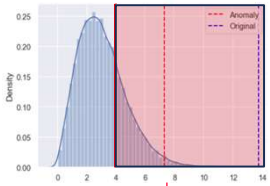
2. Test(Validation) Data 대상

- **학습 데이터와 동일한 환경 조성 [불량과 동일한 비율로 정상데이터 개수 제거]**
- **나머지 정상데이터 추가 검증 [과검 발생률 추정]** ** 과검 = 정상 품질을 불량으로 오분류

2. 모델링 과정 및 결과

데이터 전처리 # 2

3. UnderSampling 데이터 직접 설정 시도



데이터 부분 선택

1.) 추가 UnderSampling 진행

→ HARMONIC ORDER 8,24의 분포에서 임계값 기준과 어느정도 가까운 데이터만 최종 선택

→ 각 임계값 기준으로 전체 데이터의 약 50%를 추가 모델링에 활용

2.) 결과적으로 모델링 결과는 매우 저조.

→ **기존 Random UnderSampling 방법을 통한 라벨링 균형화 대비 정확도 매우 저조**

(최대 43 ~ 56%에 수렴)

→ **학습 데이터의 정보 손실을 야기하는 것으로 판단**

✓ **모델링은 Random UnderSampling 기반 전처리를 활용**

3. 모델링 과정 및 결과

모델링 결과

2. Train / Validation 결과 예시

	데이터 조합	데이터 크기	학습 정상/불량 개수	1단계 검증: 검증 정상/불량 개수	1단계 검증: 정확도(F1)	2단계 검증: 추가 정상데이터 검증 개수	2단계 검증: 추가 검증 과검률
	S32-S71-CW-H8	(4736, 20)	3733 / 55	12 / 10	**91%**	926	≈ 41%
	S32-S71-CCW-H8	(4736, 20)	3733 / 55	12 / 10	**87%**	926	≈ 47%
✓ 안정	S34-S70-CW-H8	(5426, 20)	3749 / 49	16 / 13	**76%**	1599	≈ 25%
	S34-S70-CCW-H8	(5426, 20)	4020 / 49	10 / 8	**83%**	1249	≈ 43%
	S34-S71-CW-H8	(5318, 20)	3434 / 22	12 / 10	**86%**	1840	≈ 39%
	S34-S71-CW-H8	(5318, 20)	3964 / 24	11 / 9	**80%**	1310	≈ 46%
✓ 안정	CW-H24	(20443, 12)	16235 / 119	40 / 32	**90%**	4017	≈ 20%
✓ 안정	CCW-H24	(20443, 12)	16236 / 118	40 / 32	**86%**	4017	≈ 23%

정상데이터 부분제거 대상

Ex.) 16236 → 400개로 정상데이터 제거

RandomUndersampling 이전의 기존 정확도 약 52~56%

3. 모델링 과정 및 결과

모델링 결과

3. 상관계수 전 / 후 비교

- **기존 학습 데이터의 상관관계 vs 정상/불량 불균형 비율 조정 이후(Random UnderSampling) 학습 데이터의 상관관계**

	$360_QD_CCW_COGGING_HARMONIC ORDER 24_mNm
$360_QD_CCW_COGGING_HARMONIC ORDER 24_mNm	1.000000
$230_QD_TEMPERATURE	0.065740
$240_QD_STACK_B_AREASUM	0.028527
$240_QD_STACK_A Angle	0.020659
$240_QD_STACK_A-B Skew	0.018987
$240_QD_STACK_B Angle	0.012909
$240_QD_STACK_C Angle	0.005061
$240_QD_STACK_A-C Skew	-0.004156
$230_QD_MAGN_VOLTAGE_V	-0.017803
$240_QD_STACK_B-C Skew	-0.035334
$240_QD_STACK_A_AREASUM	-0.051118
$230_QD_MAGN_CURRENT_KA	-0.093580

	$360_QD_CCW_COGGING_HARMONIC ORDER 24_mNm
$360_QD_CCW_COGGING_HARMONIC ORDER 24_mNm	1.000000
$230_QD_TEMPERATURE	0.427263
$240_QD_STACK_B_AREASUM	0.166821
$240_QD_STACK_A Angle	0.087151
$240_QD_STACK_B Angle	0.023885
$240_QD_STACK_A-B Skew	0.022563
$240_QD_STACK_C Angle	0.001025
$230_QD_MAGN_VOLTAGE_V	-0.032357
$240_QD_STACK_A-C Skew	-0.119031
$240_QD_STACK_B-C Skew	-0.212862
$240_QD_STACK_A_AREASUM	-0.344657
$230_QD_MAGN_CURRENT_KA	-0.492026

Ex.) 학습 데이터의 정상 데이터 중 일부를 무작위로 제거한 이후에도 Target과 일부 주요인자들 간의 상관관계가 어느정도 생성

꼭 높은 상관성을 지니지 않더라도 어느정도 정상/불량 판정에 영향을 끼칠 수 있는 추가적인 데이터 수집 & 추가 데이터 개수 확보의 근거 마련 가능

3. 모델링 과정 및 결과

- **모델링 결과**

4. 데이터 모델링 결과 예시

- **적용 모델 : RandomForest**

- 적용 Params : {'bootstrap': True,'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini','max_depth': None,'max_features': 'sqrt','max_leaf_nodes': None,'max_samples': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1,'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 100,'n_jobs': None,'oob_score': False,'random_state': 2021,'verbose': 0, 'warm_start': False}

- **Result (테스트셋 정확도)**

```
Confusion Matrix
[[33  7]
 [ 3 29]]

정확도: 0.861 정밀도: 0.806 재현율: 0.906 AUC: 0.866 F1: 0.853

              precision    recall  f1-score   support

          양품       0.92      0.82      0.87        40
          불량       0.81      0.91      0.85        32

    accuracy                           0.86        72
   macro avg       0.86      0.87      0.86        72
weighted avg       0.87      0.86      0.86        72
```

정상데이터 4017개 추가 검증 →

```
Confusion Matrix
[[3122  935]
 [   3   29]]
```

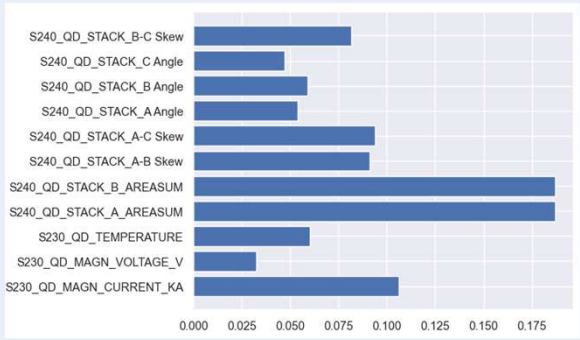
추가 정상데이터 검증에 대한 과검률 = 23%

** 정상품질에 대한 오분류 예측률= 23%

3. 모델링 과정 및 결과

- **Feature Importance 예시**

Ex.) CCW-H24



- **UnderSampling(불량 데이터의 개수와 유사한 비율로 정상데이터 제거) 과정을 걸친 학습 데이터의 상관관계**
- 대체로 균일한 변수 중요도를 띄는 것으로 확인
- 의미 있는 변수의 양품 Range 정의 및 2차 원인 분석을 토대로 실시간 품질 검증 가능성 확인

3. 모델링 과정 및 결과

LINK YOUR DATA TO VALUE

- **통계 분석 예시**

Ex.) CCW-H24

	S360_QD_CCW_CO GGING_ HARMO NIC ORDER 24_mNm	S230_QD_ MAGN_CU RRENT_K A	S230_QD _MAGN_ VOLTAG E_V	S230_QD_T EMPERATU RE	S240_QD _STACK_ A_AREAS UM	S240_QD _STACK_ B_AREAS UM	S240_QD _STACK_ A-B Skew	S240_QD _STACK_ A-C Skew	S240_QD _STACK_ A Angle	S240_QD _STACK_ B Angle	S240_QD _STACK_ C Angle	S240_QD _STACK_ B-C Skew	S360_QD _CCW_CO GGING_ HARMO NIC ORDER 12_mNm
Clean	**9.307683**	2182.388	**30.240181**	1.71E+07	1.70E+07	5.29822	10.54232	45.01802	45.01535	45.01477	5.239736	**0.0204**
Fraud	**9.150333**	2182.24	**33.18**	1.70E+07	1.71E+07	5.316133	10.5398	45.02253	45.01787	45.0158	5.219467	**0.07333**

- 최종 제품의 결과에 따른 각 조립 공정별 통계치
- 고도화 및 더 자세한 탐색을 토대로 각 조립 공정별 유의수준을 고려한 양품 탐색 가능성
- 일정 수준 이상의 차이가 발생하는 변수 (빨간색 박스 표시)들의 경우, Target과의 상관계수 & 모델(RandomForest)의 변수 중요도에서 비교적 높은 순위를 나타냄