열처리 공정 수율 개선을 위한 공정 파라미터 분석

Micro Robot

발표자: 구자윤

>> Table of contents

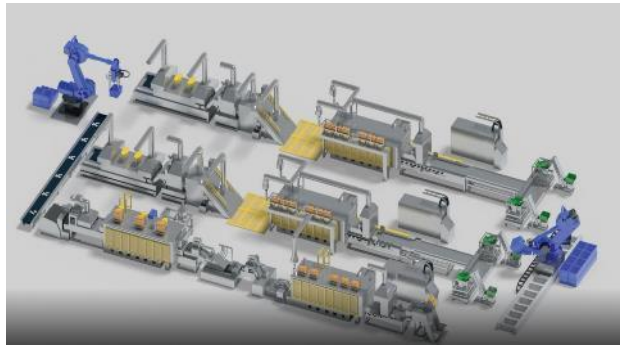
1

개요

Part 1 >> 개요

1 공정 개요

금속 부품의 열처리(Austempering) 공정으로 제품의 신율, 단면 수축률, 충격치 등의 향상으로 제품 인성을 강화하고 Crack 및 변형 감소를 위해 진행되는 공정이다.



2 이슈 사항

- 2시간 이상 소요되는 무중단 연속공정으로, 공정이 완료되어 불량 유무를 육안으로 보기 전까지 품질 상태를 확인할 수 없다.

- 현재 다양한 공정 데이터가 수집되고 있으나 품질 예측 또는 수율 분석에 활용되지 못하고 현재 설비 상태만 파악 가능하다.

3 분석 목표

- 불량 수율의 제 3사분위를 기준으로 '위험'과 '안정'으로 구분하고 이를 예측하는 분류 모델을 설계한다.

- 분류 모델을 통해 실시간 공정 상태를 '안정' 또는 '위험' 상태인지 예측한다.

- 모델이 학습한 공정 파라미터의 중요도를 분석하여, 공정 개선점을 도출한다.

Part 1 >> 개요: 모델 성능 및 공정 파라미터 분석 결과

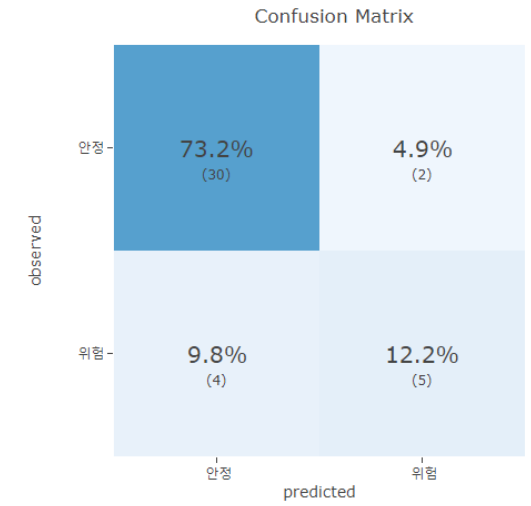
- '안정'/'위험' 분류 모델 성능 결과와 수율에 영향을 주는 공정 파라미터 중요도 분석 결과 도출 완료

- **모델 학습 결과 (Training set 70%, Test set 30%)**

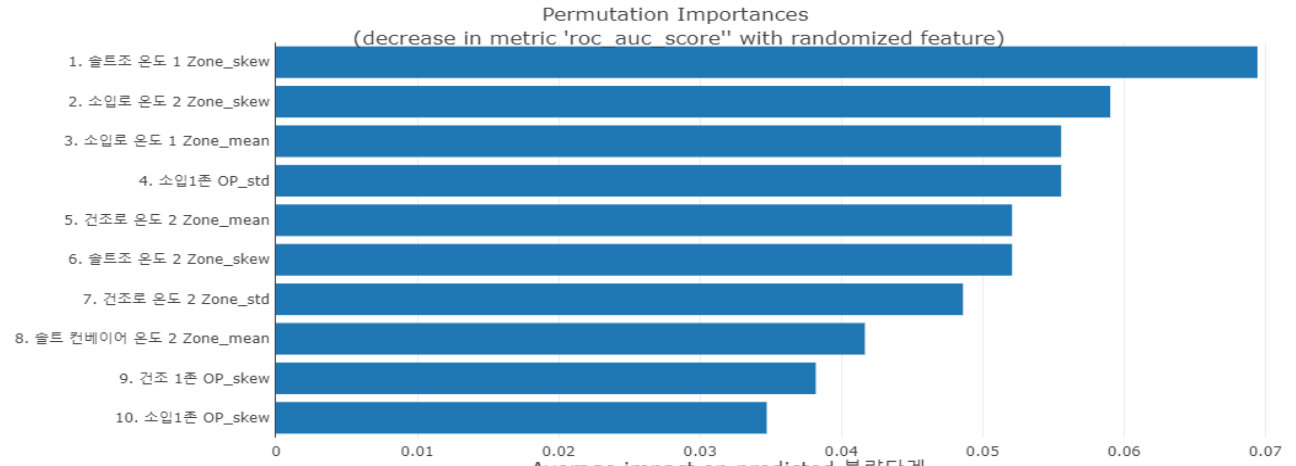
Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.7143	0.9745	1.0000	0.6364	0.7778	0.4286	0.5222
1	0.9286	0.9184	0.9286	0.9286	0.9286	0.8571	0.8571
2	0.8929	0.9847	0.9286	0.8667	0.8966	0.7857	0.7877
3	0.8148	0.9670	0.9231	0.7500	0.8276	0.6322	0.6481
4	0.8889	0.9505	1.0000	0.8235	0.9032	0.7756	0.7959
Mean	0.8479	0.9590	0.9560	0.8010	0.8667	0.6958	0.7222
Std	0.0764	0.0232	0.0359	0.1008	0.0557	0.1523	0.1212

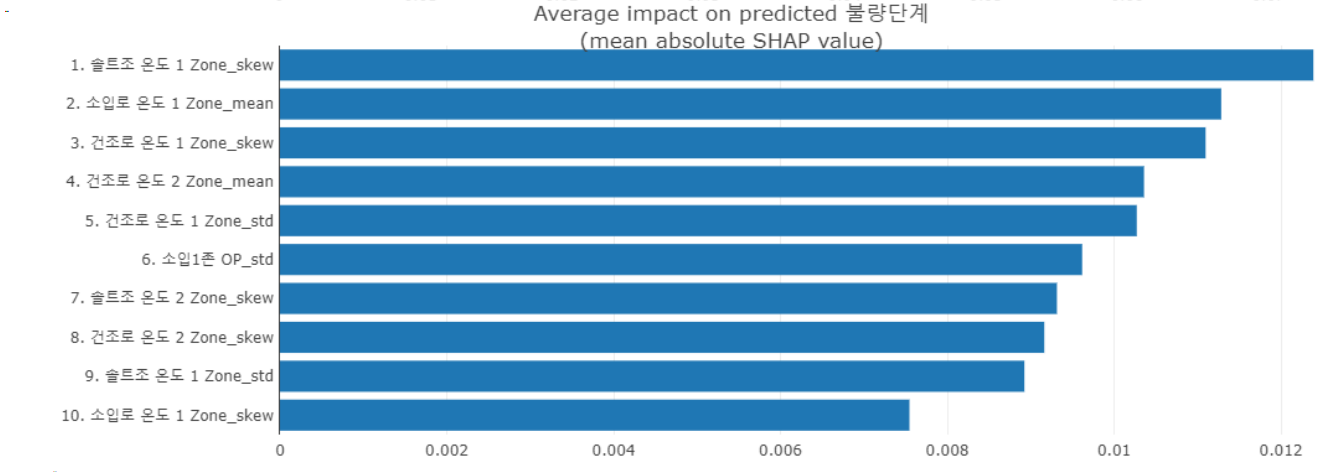
Test set

Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0.8537	0.7014	0.5556	0.7143	0.6250	0.5358	0.5424



- **공정 파라미터 중요도 결과 (Permutation & Feature importances)**





2

데이터 전처리

Part 2 >> 데이터 전처리

- 제공된 데이터는 '.csv' file로 '**공정 데이터**', '**품질 데이터**' 두가지로 나뉨
- '**공정 데이터**'는 독립 변수인 데이터가 있으며(21열 * 2,939,722행), '**품질 데이터**'는 종속 변수로 구성되어 있음
- '**파생 변수**'는 '공정 데이터'와 '품질 데이터'를 기반으로 생성된 변수이며, '소요시간', '불량률', '불량단계' 3가지를 생성하였음

종류	열 (Column)	행 (Row)	총합 (Total)	파일명 (File)
공정 데이터	21	2,939,722	61,734,162	data.csv
품질 데이터	7	136	952	Quality.xlsx
파생 변수	3	136	408	N/A

학습을 위해 준비된 데이터 세트 구성

Part 2 >> 데이터 전처리

- '**공정 데이터**'는 21개의 공정 파라미터와 2,939,722개의 공정 데이터가 존재함

종류	설명	데이터형
TAG_MIN	데이터 수집 시간 (1초 간격)	datetime
배정번호	공정 작업 지시 번호-배정번호별 생산	int
건조 1~2존 OP	각 건조 온도 유지를 위한 출력량(%)	float
건조로 온도 1~2 Zone	각 건조로 존의 온도 값	float
세정기	세정기 온도 값	float
소입1~4조 OP	각 소입 존 온도 유지를 위한 출력량(%)	float
소입로 CP 값	침탄 가스의 침탄 능력량(%)	float
소입로 온도 1~4 Zone	각 소입로 존의 온도 값	float
솔트 1~2존 OP	각 솔트존 온도 유지를 위한 출력량(%)	float
솔트 컨베이어 온도 1~2 Zone	각 솔트 컨베이어 존의 온도 값	float
솔트조 온도 1~2 Zone	각 솔트조 존의 온도 값	float

독립 변수인 '공정 데이터'의 구성 (data.csv)

Part 2 >> 데이터 전처리

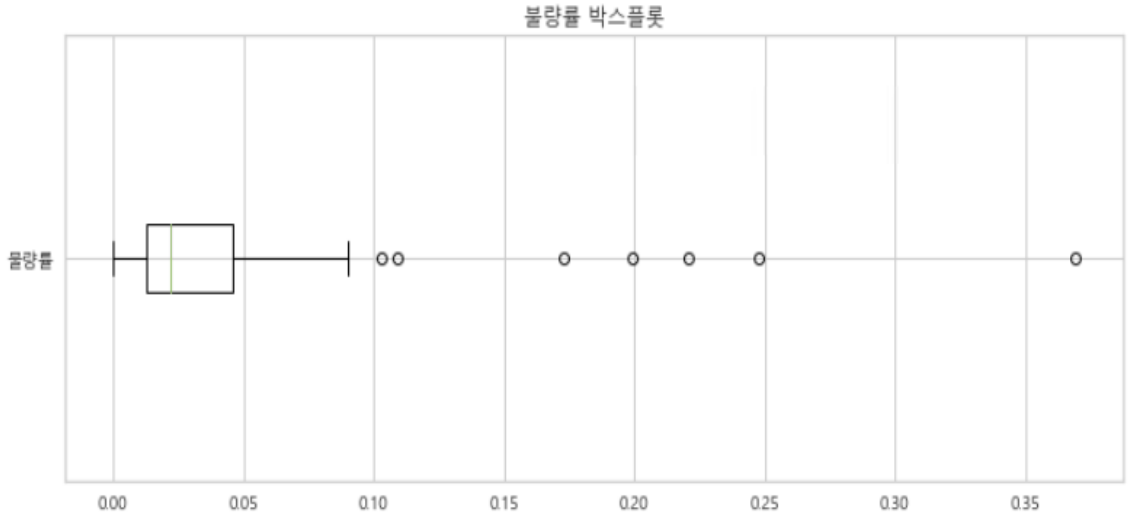
- '**품질 데이터**'는 7개의 공정 파라미터와 공정 결과에 대한 136개의 결과 데이터가 존재함
- 변수 중 '불량수량'을 '총수량'으로 나누어 '불량률'이라는 파생 변수를 생성하고, '불량률'의 제 3사분위수를 기준으로 공정의 상태를 '안정'과 '위험'으로 분류함

종류	설명	데이터형
배정번호	공정 작업 지시 번호-배정번호별 생산	int
작업일	공정 일자	datetime
공정명	진행된 공정 이름	object
설비명	작업된 설비 이름	object
양품수량	양품 생산 수량	int
불량수량	불량 발생 수량	int
총수량	전체 수량 (양품 수량 + 불량 수량)	int

종속 변수(파생 변수로 생성)를 포함하고 있는 '품질 데이터'의 구성 (quality.xlsx)

Part 2 >> 데이터 전처리

- '**품질 데이터**'는 7개의 공정 파라미터와 공정 결과에 대한 136개의 결과 데이터가 존재함
- 변수 중 '불량수량'을 '총수량'으로 나누어 '불량률'이라는 파생 변수를 생성하고, '불량률'의 제 3사분위수를 기준으로 공정의 상태를 '안정'과 '위험'으로 분류함



"불량률"의 제 3사분위수 이상은 "위험", 그 이하는 "안정"으로 분류하는 "불량단계" 변수 생성

Part 2 >> # 데이터 전처리

- 예측하고자 하는 '불량단계' 변수가 136개이므로, 독립 변수인 '**공정 데이터**'를 2,939,722개에서 136개로 변환해야 함
- '공정 데이터'의 특징을 학습 데이터로 변환하고자 기술 통계량을 추출하였으며, 데이터의 경향성을 보여줄 수 있는 '**평균**', '**표준편차**', '**왜도**'를 대표 특징으로써 사용함
- 입력된 '**공정 데이터**'와 '**품질 데이터**' 세트는 **전처리를 거쳐 Column 60, Row 136인 데이터 세트**가 됨

구분	열 (Column)	행 (Row)	생성된 파생 변수	사용 변수
공정 데이터	58	136	'소요시간'	19개 열 * 3(기술통계량)
				+ '소요시간'
품질 데이터	2	136	'불량률' (전처리 후 제거)	'총 수량'
			'불량단계'	**'불량단계' (종속 변수)**

※ 레이블이 136개로, 학습 가능한 데이터 수가 적으며 '안정'/'위험' 클래스 또한 약 3.38:1 비율로 데이터 불균형 문제 또한 존재함

Part 2 >> # 데이터 전처리

- 데이터 전처리 과정은 크게 4가지로 구성됨
- 본 데이터 세트는 학습 데이터가 부족하며, 클래스 불균형 문제를 가지고 있어 Oversampling* 과 모델 학습 시 Stratified K-Fold Cross Validation** 방법을 사용하였음

*,** 데이터 불균형 문제와 일반화 성능 향상을 위해 사용됨

Data Load

1. 데이터 세트 입력
2. 결측치 확인
3. 결측치 제거

>>

Data Preprocessing

4. 기술 통계량 확인
5. 파생 변수 생성
6. 데이터 세트 분리 (Train set/ Test set)
7. Standardization 적용
8. Oversampling 적용 (Training set)

>>

Model Training

9. Stratified 5-F CV 적용
10. Algorithms 학습 성능 비교
11. 모델 선정
12. 분류 평가 성능 지표 확인

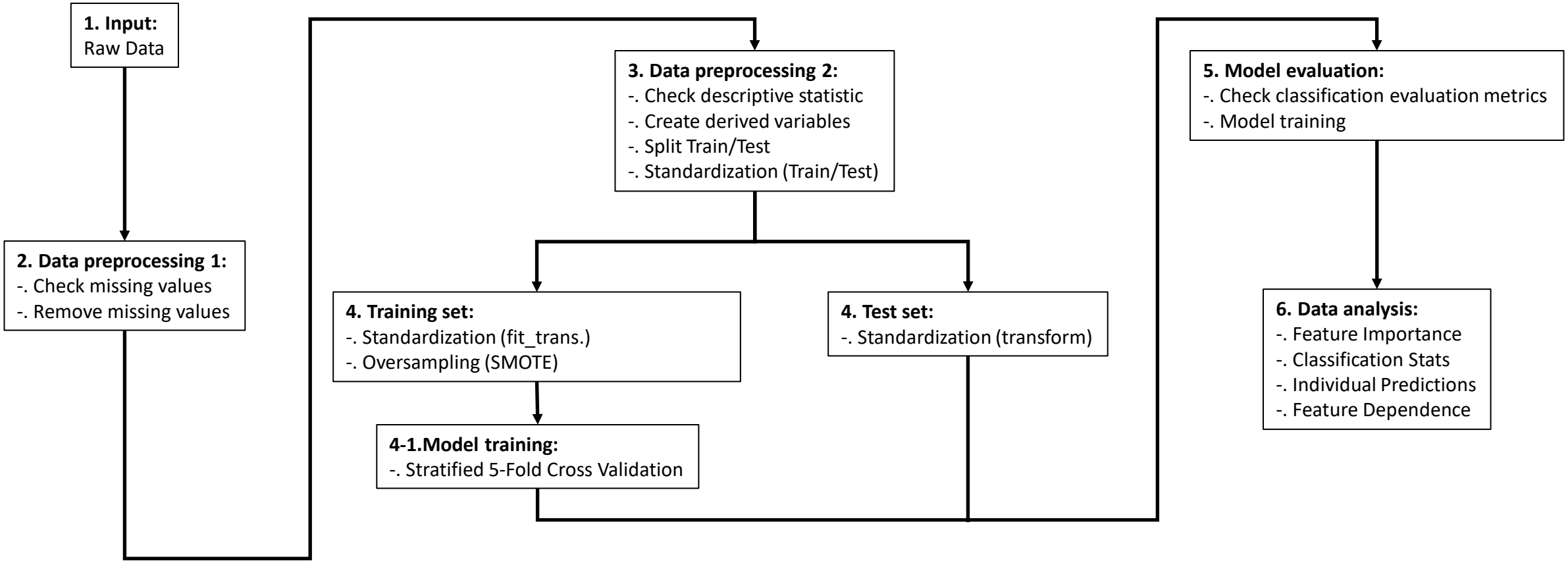
>>

Data Analysis

13. Feature Importance 분석
14. Classification stats 분석
15. Individual Predictions 분석
16. Feature Dependence 분석

Part 2 >> 데이터 전처리

- 데이터 전처리 세분화 과정은 하기 Flow Chart와 같음



3

모델 학습

Part 3 >> 모델 학습

- AutoML(PyCaret)을 이용하여 분류 모델 설계 진행
- 데이터 세트가 '**데이터 부족 문제**'와 '**클래스 불균형 문제**'를 가지고 있어 Stratified K-Fold CV 방법 적용함
- 마찬가지로, 학습 데이터 부족 문제를 해결하기 위해, Oversampling(SMOTE) 기법을 Training set에 적용함
- 모델 학습 시 이상적인 '**Oversampling Tuning**'과 '**학습 가중치 조정**' 및 **모델 성능 비교**를 위해 삼중 FOR 문을 구성하여 Fine Tuning을 진행함 (Oversampling Tuning, Model 선택, 학습 가중치 조절)
- 약 15개 Algorithms을 학습하여 **Top-3 모델**을 선정함 (RandomForest, XGBoost, **Extra Trees**)
- ①차로 Default 상태의 학습 결과를 비교하였고 ②차로 Fine Tuning 이후의 분류 성능을 비교함
- Top-3 모델을 Bagging 및 Stacking Ensemble을 적용해 보았지만, 단일 모델일 때 성능이 더 우수하였음

구분	열 (Column)	행 (Row)	Standardization 적용	SMOTE 적용 후
Training set	59	95	적용(fit_transform)	138
Test set	59	41	적용(transfrom)	N/A
계	59	136	N/A	N/A

Training set/ Test set의 구성

Part 3 >> 모델 학습

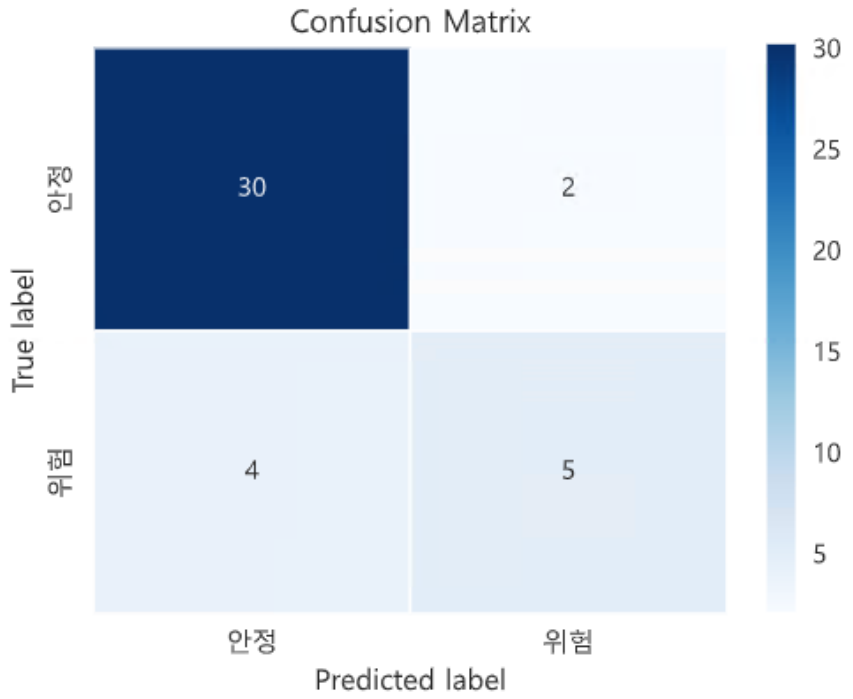
- Top-3 모델 중 Extra Trees Classifier가 빠른 연산대비 분류 성능이 우수하여 분석 모델로 채택됨
- Test set의 구성을 30%로 하였을 때 Accuracy 0.854, F1 0.625, Kappa 0.536이 도출됨
- Ensemble이 아닌, 단일 모델일 때 분류 성능이 우수하였음

학습 결과

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.7143	0.9745	1.0000	0.6364	0.7778	0.4286	0.5222
1	0.9286	0.9184	0.9286	0.9286	0.9286	0.8571	0.8571
2	0.8929	0.9847	0.9286	0.8667	0.8966	0.7857	0.7877
3	0.8148	0.9670	0.9231	0.7500	0.8276	0.6322	0.6481
4	0.8889	0.9505	1.0000	0.8235	0.9032	0.7756	0.7959
Mean	0.8479	0.9590	0.9560	0.8010	0.8667	0.6958	0.7222
Std	0.0764	0.0232	0.0359	0.1008	0.0557	0.1523	0.1212

Test set 분류 결과

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Extra Trees Classifier	0.8537	0.7014	0.5556	0.7143	0.6250	0.5358	0.5424



Extra Trees Classifier 분류 성능 지표 결과

Part 3 >> 모델 학습

- Top-3 모델 중 Extra Trees Classifier가 빠른 연산대비 분류 성능이 우수하여 분석 모델로 채택됨
- Test set의 구성을 30%로 하였을 때 Accuracy 0.854, F1 0.625, Kappa 0.536이 도출됨

Extra Trees (채택)

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.7143	0.9745	1.0000	0.6364	0.7778	0.4286	0.5222
1	0.9286	0.9184	0.9286	0.9286	0.9286	0.8571	0.8571
2	0.8929	0.9847	0.9286	0.8667	0.8966	0.7857	0.7877
3	0.8148	0.9670	0.9231	0.7500	0.8276	0.6322	0.6481
4	0.8889	0.9505	1.0000	0.8235	0.9032	0.7756	0.7959
Mean	0.8479	0.9590	0.9560	0.8010	0.8667	0.6958	0.7222
Std	0.0764	0.0232	0.0359	0.1008	0.0557	0.1523	0.1212

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Extra Trees Classifier	0.8537	0.7014	0.5556	0.7143	0.6250	0.5358	0.5424

Random Forest

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.7857	0.9337	0.9286	0.7222	0.8125	0.5714	0.5963
1	0.8571	0.8673	0.8571	0.8571	0.8571	0.7143	0.7143
2	0.9286	0.9694	0.9286	0.9286	0.9286	0.8571	0.8571
3	0.8889	0.9396	0.9231	0.8571	0.8889	0.7781	0.7802
4	0.8148	0.8846	0.7857	0.8462	0.8148	0.6301	0.6319
Mean	0.8550	0.9189	0.8846	0.8422	0.8604	0.7102	0.7160
Std	0.0510	0.0375	0.0564	0.0668	0.0444	0.1019	0.0954

Transformation Pipeline and Model Successfully Saved

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Random Forest Classifier	0.8537	0.6632	0.4444	0.8000	0.5714	0.4917	0.5227

XGBoosting

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.7143	0.8827	0.8571	0.6667	0.7500	0.4286	0.4472
1	0.8214	0.8776	0.7143	0.9091	0.8000	0.6429	0.6581
2	0.8929	0.9388	0.8571	0.9231	0.8889	0.7857	0.7877
3	0.8889	0.9341	0.8462	0.9167	0.8800	0.7769	0.7790
4	0.7407	0.8462	0.7143	0.7692	0.7407	0.4822	0.4835
Mean	0.8116	0.8958	0.7978	0.8369	0.8119	0.6232	0.6311
Std	0.0737	0.0354	0.0683	0.1025	0.0626	0.1471	0.1433

Transformation Pipeline and Model Successfully Saved

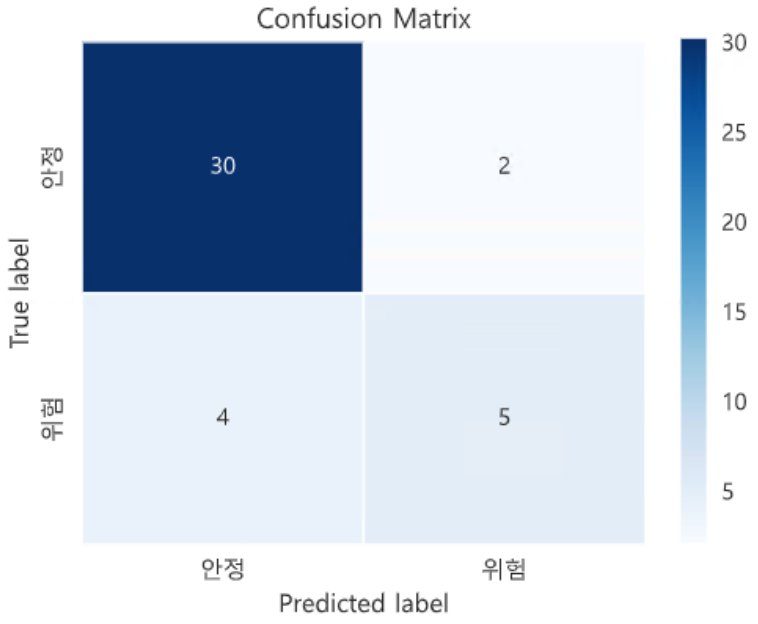
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Extreme Gradient Boosting	0.8537	0.7257	0.5556	0.7143	0.6250	0.5358	0.5424

Top-3 모델 학습 및 Test set 분류 성능 결과

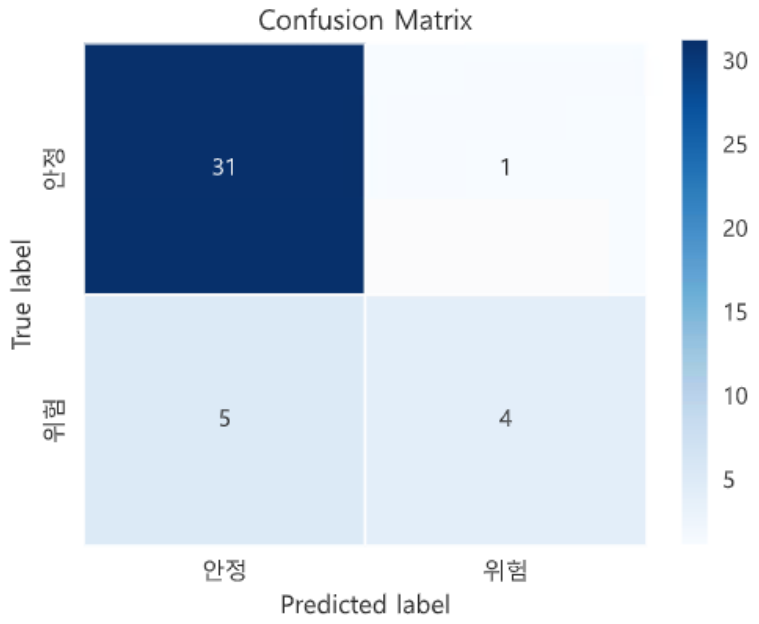
Part 3 >> 모델 학습

- Top-3 모델 중 Extra Trees Classifier가 빠른 연산대비 분류 성능이 우수하여 분석 모델로 채택됨
- Test set의 구성을 30%로 하였을 때 Accuracy 0.854, F1 0.625, Kappa 0.536이 도출됨

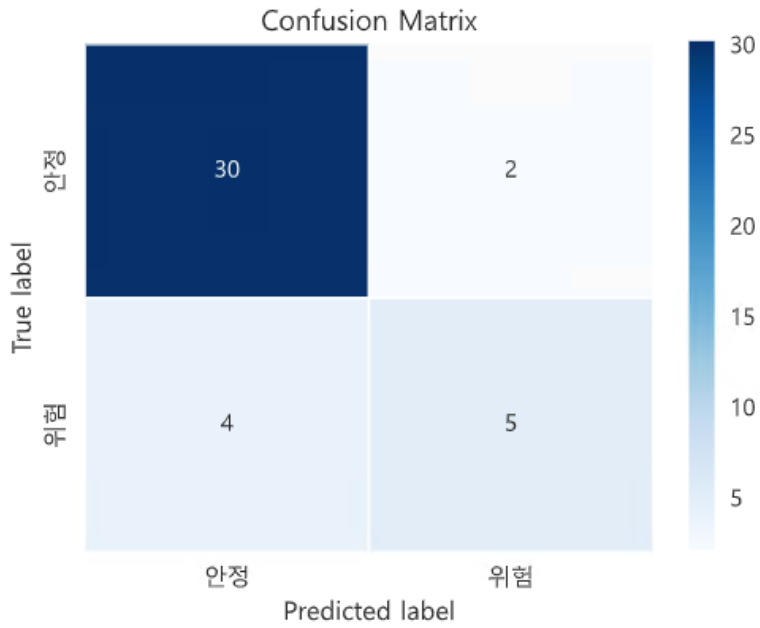
Extra Trees (채택)



Random Forest



XGBoost



Top-3 모델 Test set 분류 결과

4

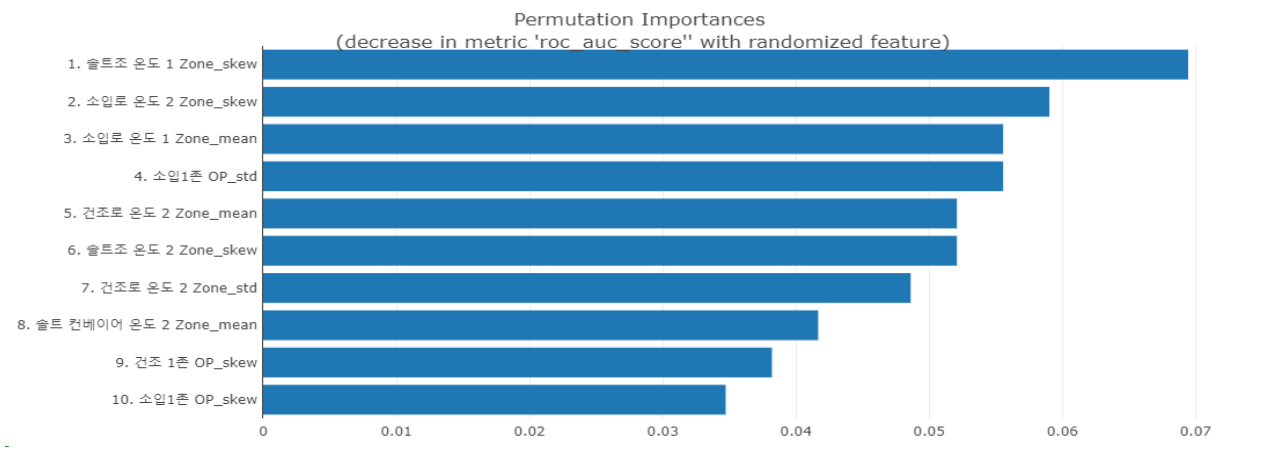
분석 및 결론

Part 4 >>

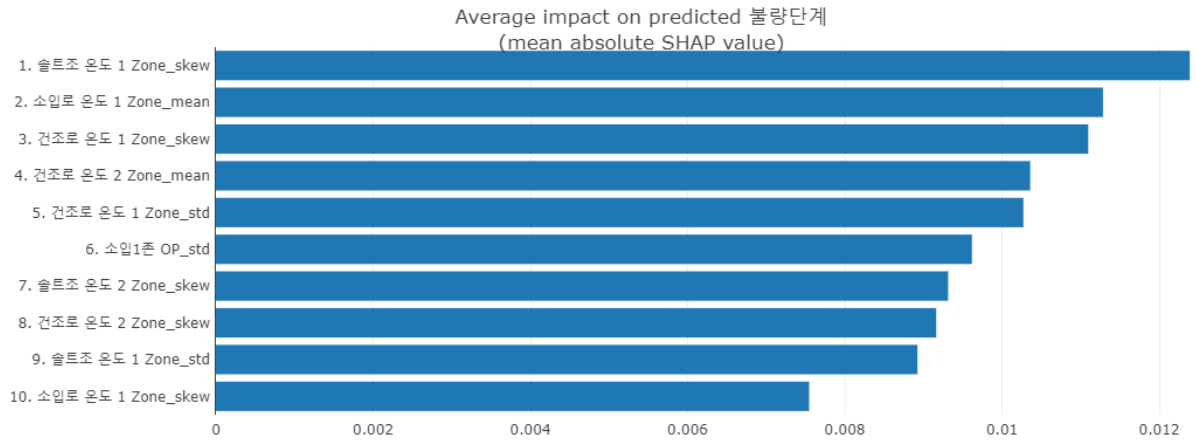
분석 및 결론: 공정 파라미터 분석

- 학습된 분류 모델의 특성 중요도 분석을 통하여 불량 수율(불량단계)에 영향을 주는 공정 파라미터 선별
- 분석 결과, '**솔트조 온도 1 Zone**', '**소입로 온도 1 Zone**', '**건조로 온도 2 Zone**'이 모델 예측에 영향을 준 중요 파라미터로 관찰됨
- 사용된 기술 통계량은 '**왜도**', '**평균**', '**표준편차**'이며 해당 특성은 '열처리 온도'가 균일하지 않고 변화하는 특성을 설명하므로 열처리 공정 엔지니어는 열처리 온도의 치우침이나 산포 관리를 통계적 분석 기법을 통해 자세히 분석해 볼 필요성이 있음

Permutation Importance* 결과



Feature Importance** 결과



* 모델이 학습 후 각 공정 파라미터를 무작위로 섞은 후 모델의 성능 변화를 관찰한 중요도
** 모델 학습 시 Tree 분할에 기여된 중요도

Part 4 >> 분석 및 결론: 공정 파라미터 분석

- **Permutation Importance** 결과와 **Feature Importance** 결과는 하기와 같음

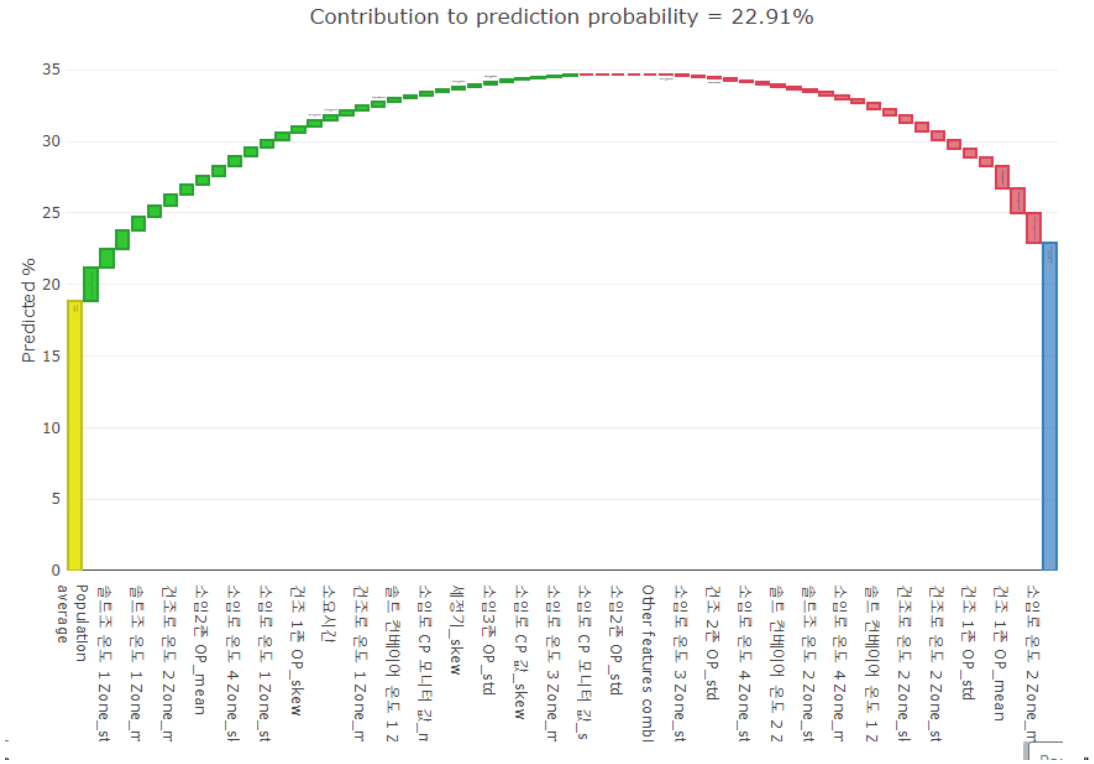
NO.	Permutation Importance	Feature Importance
1	솔트조 온도 1 Zone_skew	솔트조 온도 1 Zone_skew
2	소입로 온도 2 Zone_skew	소입로 온도 1 Zone_mean
3	소입로 온도 1 Zone_mean	건조로 온도 1 Zone_skew
4	소입1존 OP_std	건조로 온도 2 Zone_mean
5	건조로 온도 2 Zone_mean	건조로 온도 1 Zone_std
6	솔트조 온도 2 Zone_skew	소입1존 OP_std
7	건조로 온도 2 Zone_std	솔트조 온도 2 Zone_skew
8	솔트 컨베이어 온도 2 Zone_mean	건조로 온도 2 Zone_skew
9	건조 1존 OP_skew	솔트조 온도 1 Zone_std
10	소입1존 OP_skew	소입로 온도 1 Zone_skew

학습된 분석 모델의 순열 중요도 및 특성 중요도 (수율에 관련된 공정 파라미터)

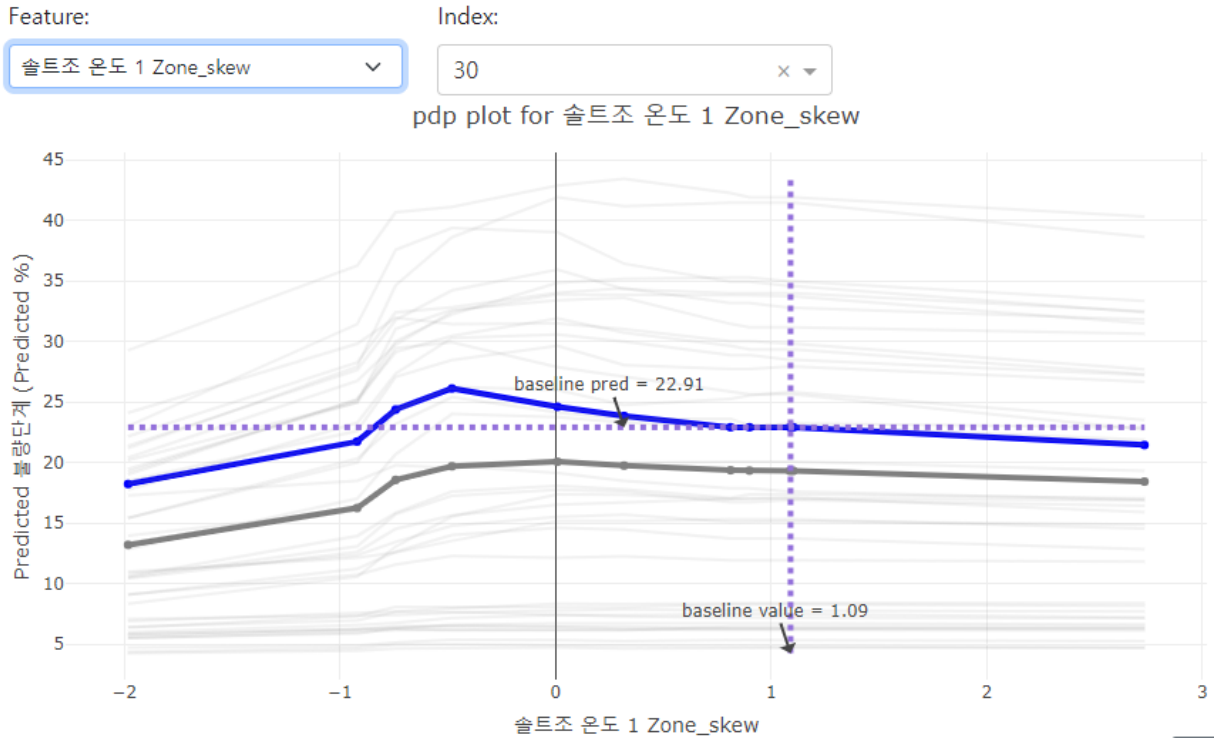
Part 4 >> 분석 및 결론: 공정 파라미터 분석

MicroRobot_분석 결과.

- 학습된 분석 모델과 Model Explainer를 기반으로 공정 파라미터와 예측 결과 간의 관계를 수치적으로 분석할 수 있음 (참고: 'MicroRobot_분석 결과.html')
- 하기 차트들은 예측 결정에 기여한 각 공정 파라미터의 기여도와 Prediction Threshold 범위를 보여줌



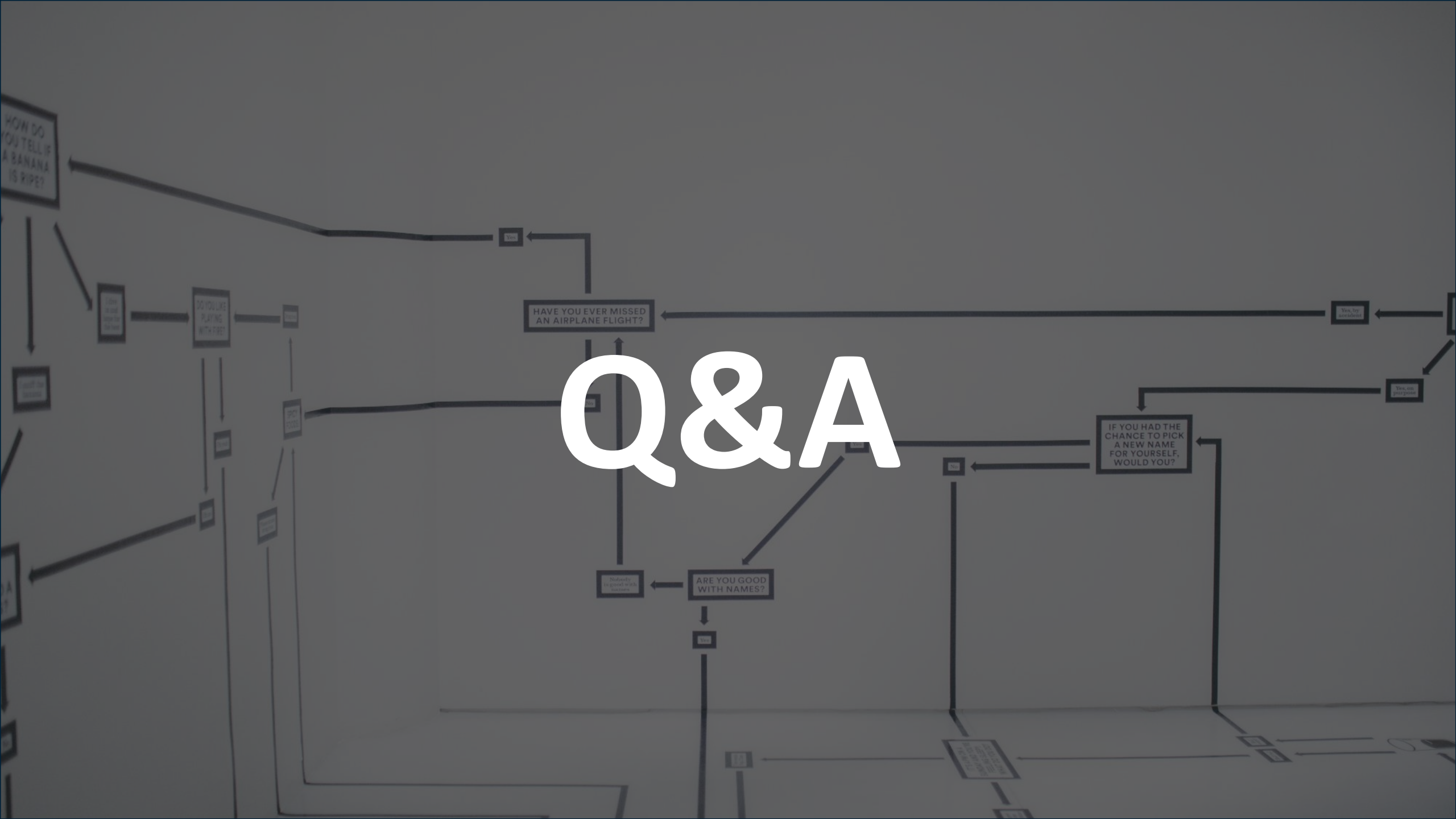
Model explainer: Contributions plot



Model explainer: Partial Dependence Plot

Part 4 >> # 분석 및 결론: 파급효과

- 본 팀은 열처리 공정의 '불량단계'(불량수율)를 예측하는 분류 모델을 설계하고, 분류 모델의 '순열 중요도' 및 '특성 중요도' 분석을 통하여, 관찰해 볼 수 있는 바람직한 공정 파라미터를 선별하였음
- 제조 라인에 상주하면서 도메인 업무를 수행하는 공정 엔지니어들이 다양한 유형의 공정 데이터 분석을 위해서 매번 모델을 설계하고 깊은 수준의 데이터 전처리를 수행하는 것이 쉽지 않음
- 따라서, 본 팀은 기존 도메인 업무와 병행 가능한 수준에서 쉽고 효과적으로 기계학습을 이용한 공정 데이터 분석을 목표로 프로젝트를 수행하였고, 도메인 지식이 없는 상태와 학습 데이터가 부족한 상황에서도 공정 상태를 '안정' 또는 '위험'으로 예측하는 모델 설계를 완료함
- 위 과정을 통하여, 실무 공정 엔지니어는 공정 내 중요 파라미터를 선별하고, Six Sigma와 같은 통계적 품질 관리와 데이터 분석을 통하여 공정 수율 개선을 위한 설비 개조와 분석 파라미터 발굴 등의 Advanced 업무를 수행해볼 수 있음
- 이와 비슷한 H/W 데이터와 품질 계측 데이터 수집이 가능한 중소제조기업에서는, 위와 같은 간단한 데이터 처리와 기계학습 과정만으로 공정 데이터 분석이 가능할 것으로 생각됨


HOW DO YOU TELL IF A BANANA IS RIPE?
Yes
HAVE YOU EVER MISSED AN AIRPLANE FLIGHT?
Yes, by accident
Yes, on purpose
DO YOU LIKE PLAYING WITH FIRE?
IF YOU HAD THE CHANCE TO PICK A NEW NAME FOR YOURSELF, WOULD YOU?
No
Nobody is good with names
ARE YOU GOOD WITH NAMES?
Yes

Q&A

Thank you