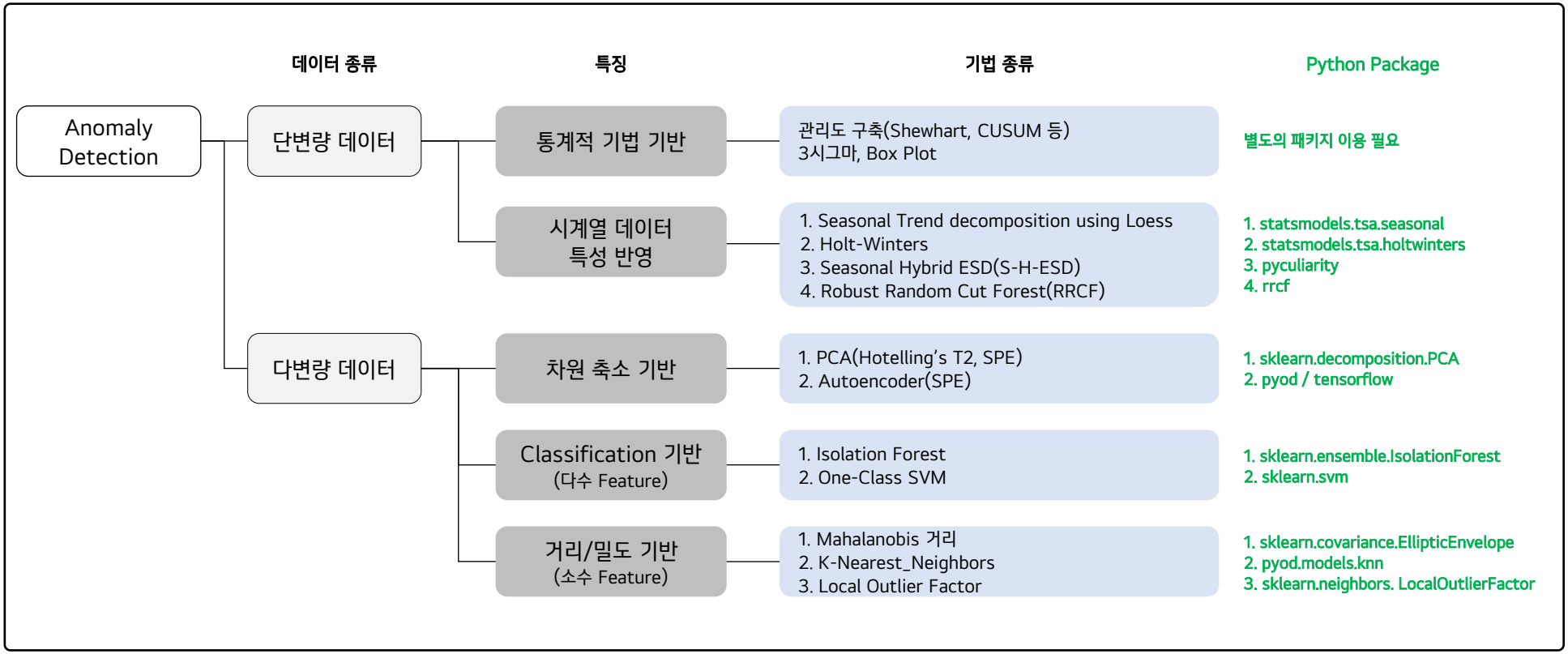
Anomaly Detection Algorithm Guide

데이터 및 문제 해결 상황에 맞는 이상 탐지(Anomaly Detection) 알고리즘 선택가이드임.

모든 데이터에 완벽한 이상 탐지 알고리즘은 없으며 알고리즘 별 주요 Hyper parameter 설정을 통해 데이터에 Fit시켜야 함.

선택 가이드 구성도



Anomaly Detection Algorithm Guide

선택 가이드 구성도

기법	장점	단점
3 sigma	• 구현이 쉽다. • 데이터 분포가 변하지 않는 이상 재학습이 필요 없음	• 정규 분포가 가정되어야 함. • Feature간 상관관계 파악이 어려움 • 데이터의 양이 충분하지 않다면, Outlier가 통계치에 영향을 미칠 수도 있음.
Box Plot	• 데이터를 눈으로 확인하기 어려울 때, 그림을 이용해 데이터의 범위를 빠르게 파악할 수 있음.	• 가운데 선은 평균이 아님. • 오해 소지가 있음(Median과 Mean은 다름.)
STL	• 시계열 데이터에 대해 분기별, 월별, 일별 분해 모두 가능 • MA(Moving Avg) 방식이 아니기 때문에 데이터 유실 없음. • 돌발스런 이상치에 대해 추세, 주기에 영향을 미치지 않음.	• 시간 데이터 전처리 필수적임. • 덧셈 분해 기능만 제공
Holt-winter	• 연산량이 적음, 큰 데이터 세트에 대해서 리소스 절약, 자동화 가능	• 변량 데이터에 대해서만 적용 가능, 상관관계 고려 X • 계절성이 없는 데이터에 대해서는 성능 저조 • 변동이 적은 계절성 데이터에 대해서는 민감하게 탐지할 우려
S-H-ESD	• 데이터 내의 노이즈에 어느 정도 대응할 수 있음. • 급작스런 상승하는 이상치를 탐지할 수 있음	• 이상 탐지가 안 되는 경우가 아래와 같이 존재함. - 점진적 증가 신호(seasonal grow) - 점진적 증가하는 신호에서의 음의방향 이상치 (Negative seasonal anomaly) - 평면적 신호 (Flat signal)
RRCF	• S-H-ESD에서 탐지되지 않았던 경우에 어느 정도 대응 가능함. • Batch 및 Streaming data 모두 활용 가능함. • Subsampling을 통한 적은 연산량	• 분리를 위한 선을 수직과 수평으로만 자르기 때문에 잘못된 scoring이 발생할 수 있음
PCA-Hotellings'T2	• 고차원에 데이터의 특징(잠재변수)을 추출할 수 있음. • 선택한 변수들의 해석이 용이함. • PCA의 잠재변수 축을 이용하여 거리기반의 이상탐지를 실시하기 때문에 직관적임.	• Threshold에 따라 민감함.
Autoencoder Reconstruction	• 데이터 Label이 존재하지 않아도 사용 가능 • 고차원에 데이터의 특징(잠재변수)을 추출할 수 있음. • Auto encoder를 기반으로 다양한 알고리즘 존재	• Hyper parameter (※ hidden layer) 설정이 어려움. • Loss(Reconstruction Error)에 대한 threshold 설정이 어려움.

Anomaly Detection Algorithm Guide

선택 가이드 구성도

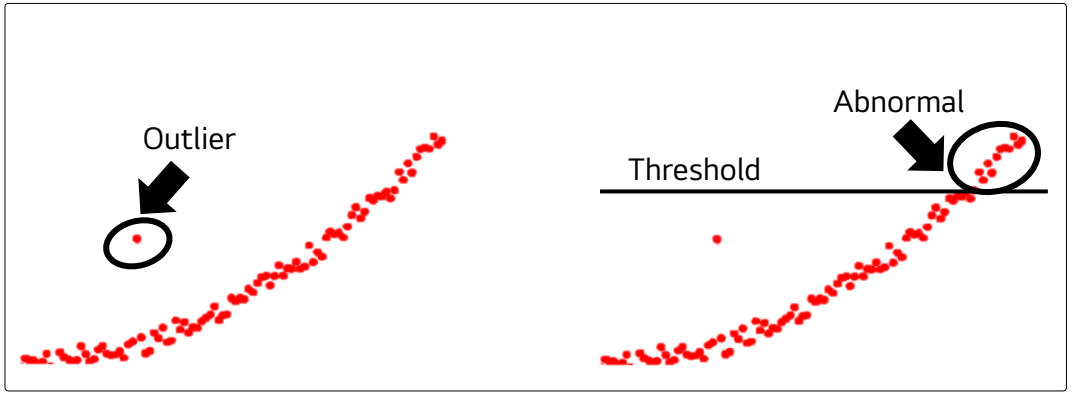
기법	장점	단점
Isolation Forest	• 군집기반 이상탐지 알고리즘에 비해 계산량이 매우 적음. (※ Sampling 사용 Tree 생성) • Anomaly Detection 성능 우수 • Train Data에 이상치가 없어도 Test Data에서 우수한 성능을 보임.	• 분리를 위한 선을 수직과 수평으로만 자르기 때문에 잘못된 scoring이 발생할 수 있음 (※ 대안 방법: Extended Isolation Forest)
One-Class SVM	• 데이터 Label이 존재하지 않아도 사용 가능 • 저차원이나 고차원의 적은 데이터에서 일반화 능력이 좋음 (데이터 특성이 적어도 성능이 좋게 나오는 편, Robust 함)	• Kernel 기반의 방법론, 데이터가 늘어날 수록 연산량이 크게 증가함. • Scaling과 Hyper parameter에 민감함.
Mahalanobis 거리	• 비선형 관계의 데이터에 활용 가능 • 데이터에 자체에 대한 가정이 필요 없음	• 변수 간의 관계가 모두 독립이라면 유클리드 거리와 같은 개념 • 변수 간의 상관성이 명확이 알려져 있지 않은 경우 적용하기 어려움
K-Nearest_Neighbors	• 단순하고 효율적임 • 기존 분류 체계 값을 모두 검사하여 비교하므로 높은 정확도를 보임. • 수치 기반 데이터 분류 작업에서 성능 우수함. • 기존 데이터를 기반으로 하기 때문에 데이터에 대한 가정이 없음.	• 기존의 모든 데이터를 비교해야 하기 때문에 데이터가 많으면 많을 수록 처리 시간이 증가 • 특징과 클래스간 관계를 이해하는데 제한적 • 카테고리컬 데이터를 위한 추가 처리가 필요
Local Outlier Factor	• 굉장히 밀집한 클러스터에서 조금만 떨어져 있어도 이상치로 탐지 • KNN과 다르게 특별한 라벨링이 없어도 사용할 수 있음 • Local Outlier를 탐지할 수 있음	• 데이터의 차원수가 증가할 수록 연산량 증가 • 이상치 판단 기준 설정 어려움 (밀집도가 다른 여러 클러스터가 존재한다면 민감하게 반응함)

Outlier vs Abnormal vs Novelty

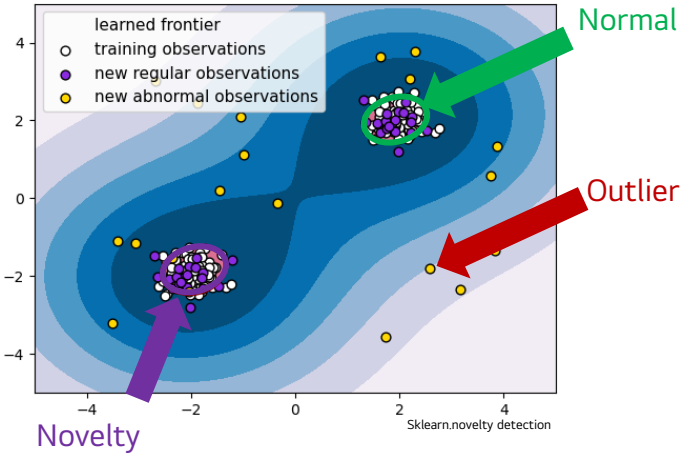
이상 탐지를 학습 데이터에 따라 아래와 같이 나눠 볼 수 있으며 학습할 데이터를 어떻게 정의하는가에 따라서 문제의 성격과 해결 방법이 다름.
새로운 관측치가 기존 분포에 속하는지 기존 분포를 벗어나는지 구분함

- Outlier: 관측된 데이터의 범위에서 많이 벗어난 아주 작은 값이나 큰 값으로 분석 결과 해석 시 오해를 발생시킬 수 있기 때문에 사전 제거 필요
- Abnormal: Domain-Knowledge 기반의 문제 상황 범주에 속한 데이터
- Novelty: 학습된 데이터 외의 새로운 패턴의 데이터

Outlier vs Abnormal



Outlier vs Abnormal vs Novelty



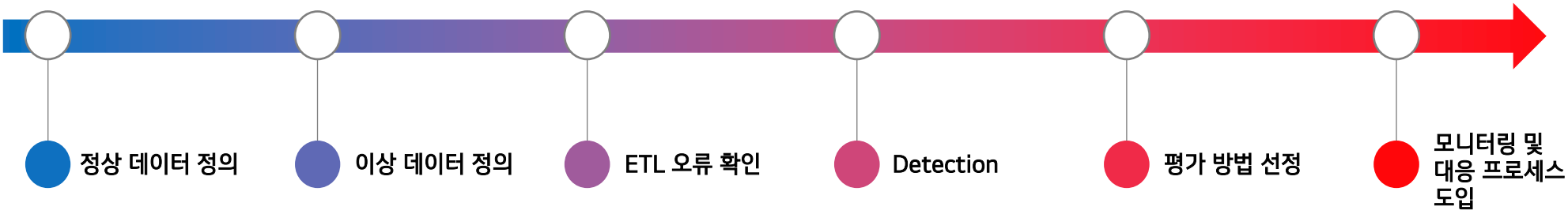
Outlier/Abnormal Detection

학습 데이터를 통해 정상 데이터의 범위를 결정하고, 이를 초과할 경우 Abnormal로 간주
훈련 데이터 셋에 정상 샘플과 이상치 샘플을 모두 포함하고 있음.

Novelty Detection

학습 데이터 내에 존재하지 않는 패턴의 데이터를 탐지
탐지 목표 1: In-distribution Test Set의 정확한 예측
탐지 목표 2: Out-of-distribution 데이터 셋은 걸러내는 것

Anomaly Detection Process (1/2)



정상 데이터란,
Domain Knowledge 기반으로 특별한 이상이 없는 데이터

혹은

이상으로 정의된 데이터 외의 데이터

이상 데이터란,
Domain Knowledge 기반으로 이상으로 정의된 데이터

혹은

사전에 감지하고 있는 데이터

정합성 검증 필요
(수집/추출 과정에서 오류가 없는가)

데이터 ETL 과정
Extract Transform Load 별 이슈 확인
예) 센서 고장값 확인
예) 결측치 확인

데이터 타입, Label 유무, Train Data에 따라 하기 표와 같이 나눌 수 있음.

데이터에 적절한 방법으로 Detection 수행

평가방법 선정
- 결과를 평가하기 위한 Threshold 설정 (현장 협의 필요)

1. 현업 모니터링 설계
2. 이상 감지 후 점검하기 위한 Action 프로세스 설계
3. False Alarm 이벤트에 대한 대응체계 수립
4. 재학습 체계 수립

Anomaly Detection의 종류

학습 유형	패턴 유형	데이터 유형	데이터 종류
지도 학습 (Supervised)	이상 탐지 (Outlier/Abnormal Detection)	단변량 데이터 (Univariate)	시계열 데이터 (Time-Series)
반 지도 학습 (SemiSupervised)	신규성 이상 탐지 (Novelty Detection)	다변량 데이터 (Multivariate)	비 시계열 데이터 (Non Time- Series)
비지도 학습 (Unsupervised)			

예시 프로세스 – 설비 이상 탐지

센서 데이터 수집
데이터 수집 방법 결정 및 수집
실시간 센서 데이터 수집 체계 구축필요

이상 증상 정의
고장 나기 전 이상 증상에 대한 정의
현장 미팅을 통해 사전 이상 증상을 정의

예측 모델링
이상 증상을 예측하기 위한
Classification 분류 모델링 이용

성능 평가 및 현장 적용
중요 인자 관리용 모니터링 시스템 구축
고장이 났을 때의 프로시저를 구축

Anomaly Detection Process (2/2)

예시 프로세스 – 품질 이상 탐지 모니터링

❶ 데이터 수집

- 데이터 수집 방법(I/F) 결정 및 수집

① 공정 센서 데이터가 수집되고 있는 원천 DB 확인

② 데이터 전송 방식(SFTP, FTP)을 통해 분석 Server로 데이터 수집

③ 특정 주기 단위로 데이터 실시간 수집

- 품질 데이터 적재 DB 확인 및 실시간 수집 시스템 구축
- 실시간 수집 체계 구축 전 Dump 형태로 데이터 전달 받은 후 분석 진행

❷ 모델링

- 머신러닝 알고리즘 활용 예측 모델링

- 생산품의 무게나 성능 같은 연속형 Target을 예측할 때는 회귀 계열 알고리즘을 사용
- 양품과 불량품을 예측할 때는 분류(Classification) 계열 알고리즘 사용

❸ 중요인자 도출

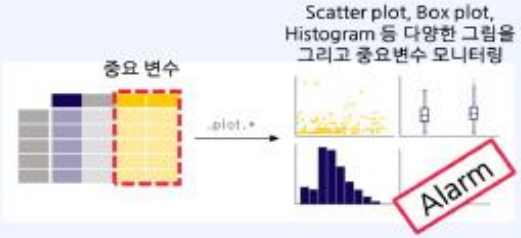
- Feature IMP 분석 및 중요인자 도출

다양한 공정 변수의 Feature IMP 비교

- 모델링 완료 후 회귀 계수, Feature IMP 활용 중요인자 도출
- 중요인자 Scatter plot을 통한 모델 설명력 확보

❹ 모니터링 시스템

- 중요인자 관리를 위한 모니터링 시스템 구축



- 품질 중요인자 관리를 위한 공정 변수 모니터링 시스템 개발
- 중요변수가 특정 수치를 넘어갈 시 경고를 전송하여 점검할 수 있도록 운영 체계 구축

기대 효과 산출

주요 공정 변수 모니터링 전/후 생산량 비교

주요 공정 변수 모니터링 전/후 불량률 비교

주요 공정 변수 모니터링 전/후 폐기 비용 비교

Anomaly Detection Project

프로젝트 이해관계를 알고 관계 간 적절한 관리가 필요함.

널리 알려진 시행착오와 프로젝트 진행을 지연/방해하게 되는 요소들을 이해하고 사전에 해당 내용을 점검, 프로젝트의 불확실성을 줄여야 함.

프로젝트 일정 수립 시 머신러닝 프로젝트의 불확실성*을 고려하여 범위의 확정, 변경되지 않는 기획이 필요함.

머신러닝 프로젝트의 불확실성: 새로운 아이디어, 반복적인 실험, 예측할 수 없는 결과로 가시적인 결과를 예상하기 어려움

프로젝트 이해관계	비즈니스 관점의 점검 사항	데이터 관점의 점검 사항
1. 이상 탐지 필요 부서(제조 공정 현장)	1. 머신러닝 문제 정의	1. 데이터의 위치/권한 확인
2. 데이터 사이언스/TF팀	2. 요구사항 점검(KPI, 예측 주기 등)	2. 이상 감지 여부 확인 - 이상 데이터 포함 여부/발생주기 - 이상에 영향을 주는 Feature 유무
3. IT/외주개발	3. 가용 가능한 자원/환경 점검	
4. CEO/임원	4. 선행 모델의 장/단점 파악	

프로젝트 일정 분배 예시

● 문제 정의/구체화 — 2M
- 목표 설정,지표 결정
- 제약 조건 확인
- 평가 방법 설계

⇩

● 데이터 훑어보기 — 0.5 ~ 1M
- 데이터취득
- 데이터이해

⇩

● 베이스라인 설계 — 0.5 ~ 1M (변경 가능)
- 베이스라인모델구현
- AutoML 활용
- 서베이 논문 조사

⇩

● 성능 개선 실험 — 4M (프로젝트 일정에 맞추어 변경할 것)
- 성능 개선 아이디어
- 제시 및 반복적인
- 실험 결과 확인

⇩

● 최종 보고 / 배포

<Do List>

잦은 보고를 통한 현재 상태 공유 및 방향성 합의

일정에 마진 포함

공유하는 성능에도 마진을 포함