공학전문석사 학위 연구보고서

기계학습 모델 기반의 저밀도 폴리에틸렌 용융지수 예측 연구

LDPE melt flow index prediction study based on machine learning model

2023년 2월

서울대학교 공학전문대학원
응용공학과 응용공학전공
김항래

기계학습 모델 기반의 저밀도 폴리에틸렌 용융지수 예측 연구

LDPE melt flow index prediction study based on machine learning model

지도교수 구윤모

이 프로젝트 연구보고서를 공학전문석사 학위 연구보고서로 제출함

2023년 2월

서울대학교 공학전문대학원

응용공학과 응용공학전공

김항래

김항래의 공학전문석사 학위 연구보고서를 인준함

2023년 2월

위원장 _____ 곽 우 영 (인)

위　원 _____ 서 은 석 (인)

위　원 _____ 구 윤 모 (인)

초 록

저밀도 폴리에틸렌 용융지수 조절은 공정 샘플 분석 결과를 기반으로 운전원이 관리 방향을 결정하고 공정 운전 조건에 반영하여 실행된다. 그러나 샘플 분석이 없는 시간대는 운전원이 공정 상황 분석 후 공정 간이 용융지수 측정값을 참조해 용융지수를 예측 조절하며 결과적으로 이에 따라 전체 제품 물성 산포가 결정된다. 본 연구는 운전원 숙련도에 따라 발생하는 제품 품질 편차를 개선하기 위해 숙련도와 무관한 공정 데이터를 활용한 기계학습 모델 기반의 용융지수 예측 연구를 수행했다.

공정에서 수집중인 온도, 압력 데이터를 독립변수로 사용하였고 실험실에서 분석 및 기록되는 용융지수는 예측 및 분류모델에 활용하기 위해 각각 연속형, 이산형 두 가지 형태로 전처리 하였다. 연속형 데이터를 사용한 다중 회귀 분석의 예측 성능은 R^2, RMSE, MAE를 평가지표로, 이산형 데이터를 사용한 XGBoost등의 분류 모델은 정확도, 정밀도, 재현율을 평가지표로 하여 현재 사용중인 간이 용융지수 분석기 대비 성능 개선율을 비교하였다. 또한 ARIMA, LSTM의 시계열 분석 도구로 용융지수를 예측하여 딥러닝 기술의 활용 가능성을 검증했다.

본 연구는 대표적 석유화학 공정에서 예측, 분류, 시계열 방법론의 유용성을 검증했으며, 용융지수를 주요 물성으로 간주하는 다양한 석유화학공정에 기계학습 방법론을 제시한 것에 연구 의의가 있다.

주요어 : 저밀도폴리에틸렌, 기계학습, 용융지수, 물성예측, 품질관리

학번 : 2021-21941

목 차

그 림 목 차

표 목 차

제 1 장

서론

1.1 연구 배경

우리나라 석유화학산업은 1970년대 정부 주도로 육성된 이래 2020년 현재 에틸렌 생산 능력 기준 세계 4위 규모로 성장했다. 하지만 설비 노후화 및 숙련된 인력의 정년퇴직으로 그 동안 축적된 경험과 기술 유출이 우려되는 상황이다[1]. 이에 대비한 설비 투자 및 인적 기술 시스템화가 진행중이며 그 일환으로 공정 데이터를 활용한 디지털 혁신을 추진중이나 관련 연구 사례는 아직 미진하다.

저밀도 폴리에틸렌(low density poly ethylene, LDPE)은 대표적 석유화학 제품으로 온도, 압력 등의 공정 조건에 따라 제품 물성이 결정된다. 제품 물성 중 하나인 용융지수(melt flow index, MI) 산포 조정은 공정에서 샘플링한 제품의 실험실 분석 결과를 확인하여 정해진 조정 범위 안에서 운전원이 실시한다. 그러나 샘플 분석이 없는 시점은 공정에 설치된 간이 용융 지수(automatic melt index, AMI) 및 운전원의 경험과 역량으로 물성을 예측하여 공정 조건을 조정한다. 샘플링 빈도를 늘릴 경우 운전원 영향을 최소화시켜 MI를 안정적으로 유지할 수 있으나 품질 비용은 증가한다. 그러나 공정 변수로 기계학습 모델을 만들수 있다면 추가되는 품질 비용 없이 MI를 실시간 예측할 수 있다.

석유화학공장은 공정에 설치된 수백 수천가지의 센서 정보를 실시간으로 수집하여 제어, 모니터링 사용하고 있다. 무엇보다 서버에 저장된

방대한 데이터는 기계 학습에 그대로 활용 가능하다. 따라서 이러한 제반 조건을 활용해 제품 물성 예측 기계 학습 모델 연구를 진행하고자 한다.

1.2 연구 목적

LDPE MI 예측 방법을 운전원 개인 경험 기반 방식에서 실시간 공정 데이터를 활용한 기계 학습 모델로 예측할 수 있도록 연구를 수행했다. 기계 학습으로 예측한 물성이 실제 공정 운전에 활용될 경우 객관적 데이터에 기반한 공정 조건 조정으로 현재 대비 균일 물성의 고품질 제품 생산이 가능하여 제품 시장 경쟁력이 향상된다. 이를 위해 공정에서 수집중인 온도, 압력의 다양한 센싱 데이터로 구현한 기계 학습 모델과 실험실에서 분석한 값을 비교하여 기계학습 모델의 유용성을 검증하였다.

1.3 연구 보고서 구성

본 연구 보고서는 총 5 장이며 각 장은 다음과 같이 구성되어 있다.

제1장은 연구 배경 및 목적 그리고 보고서 구성에 대해 정리했다.

제2장은 기계학습 독립변수와 관련된 LDPE 제조 공정 및 종속변수인 LDPE 주요 물성에 대해 요약하였으며 마지막절에서는 본 연구와 관련된 기계학습 이론과 모델을 설명하였다.

제3장은 제조 공정에서 연구 대상을 한정하는 과정과 구체적 연구 진행 방법을 요약하였고 연구 성과를 평가한 지표에 대해 기술하였다.

제4장은 기계 학습 모델 구현 결과를 구체적으로 기술하였다. 종속변수 형태에 따라 연속형, 범주형 데이터 모델로 구분하여 결과를 정리했으며 마지막 절은 시계열 모델로 구현한 기계 학습 모델에 대해 설명했다.

제5장은 결론으로 본 연구 성과를 요약 및 고찰하였다.

제 2 장

이론적 배경

2.1 LDPE 제조 공정

LDPE 제조공정은 압축, 반응, 분리, 압출 및 제립, 저장 및 포장 5 단계로 구분된다. 기계학습에 사용된 독립변수는 공정에 설치된 주요 센서 측정값을 사용했다. 제품 물성은 전 제조 공정의 영향을 받으나 주된 영향을 주는 공정은 반응, 분리, 압출 및 제립 과정이다.

LDPE 제조는 주원료인 에틸렌을 압축기로 반응 압력까지 압축하여 시작된다. 압축된 에틸렌과 개시제를 화학 반응 온도까지 승온한 반응기에 공급하면 중합 반응이 일어나 20%의 LDPE가 생성된다. 이후 분리기에서 미반응된 에틸렌은 1차 압축기로 재공급되고 LDPE는 압출기를 거쳐 최종 저장 및 포장 공정으로 이송된다.

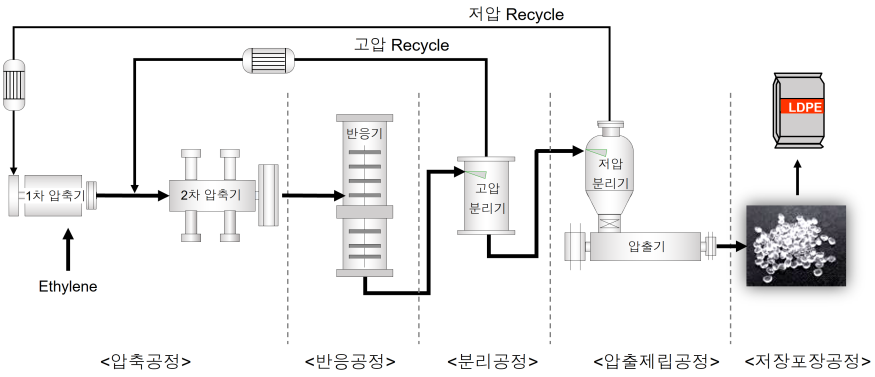

그림 1: 제조 공정

2.1.1 압축 공정

NCC에서 원유를 정제하여 얻은 납사(naphta)를 분해하면 에틸렌이 만들어진다. 에틸렌은 30bar 압력으로 이송 배관을 통해 LDPE 공장에 실시간으로 공급된다. 공급받은 에틸렌은 1차 압축기에서 10배로 압축하여 300bar까지 승압한 이후 2차 압축기에서 최종 압력까지 제품 용도에 따라 승압한다. 저분자량 제품은 통상 1,000bar까지 에틸렌을 승압하며 고분자량 제품은 최대 2,000bar까지 압력을 올려 반응 공정에 투입한다.

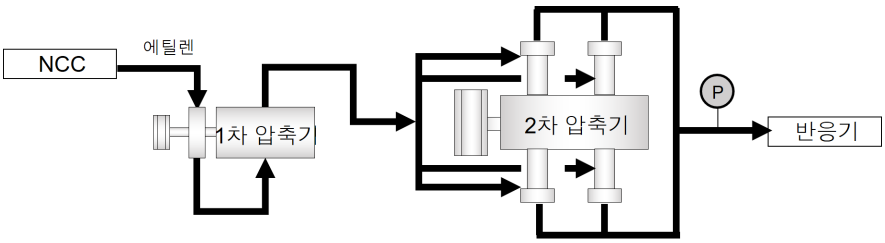

그림 2: 압축 공정

2.1.2 중합 공정

Autoclave 반응기 내부는 여러 개의 구역으로 나뉘어져 있으며 각 구역 온도 프로파일(profile)에 따라 제품 물성이 최종 결정된다. 각 구역의 온도는 개시제 공급량으로 조절하며 설정된 온도에 따라 개시제 공급량은 자동으로 조절된다. 제품 물성 조정이 필요한 경우 운전원은 1초 단위로 모니터링 되는 반응기 압력과 반응기 구역별 설정 온도를 미세 조정하여 목표 물성을 맞추게 된다. 최종 반응기에서는 20%의 에틸렌이 고분자 합성 수지(LDPE)로 전환되며 용융상태의 합성수지는 나머지 미반응 에틸렌과 후단의 분리기로 이송된다.

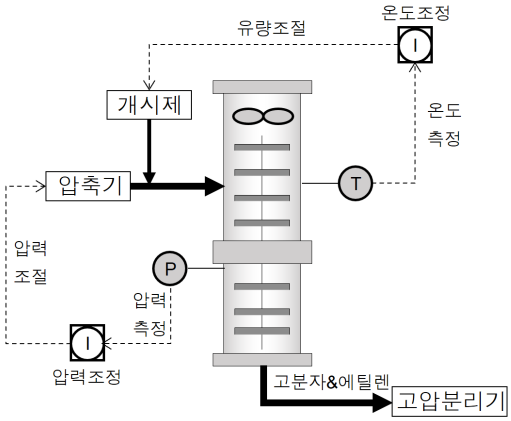

그림 3: 중합 공정

2.1.3 분리 공정

분리기로 이송된 미반응 에틸렌에는 아직 활성이 남아있는 개시제가 소량 존재해 저분자량 합성수지 및 왁스를 생성하게 된다. 분리기에 냉각 에틸렌을 주입해 내부 온도를 낮춰 개시제 활성을 제거하며 미반응 에틸렌은 공정에 재순환시키기 위해 상단 출구에서 압축기로 보내진다. 하단에서는 고분자 합성 수지만 분리되어 저압분리기로 이송된다.

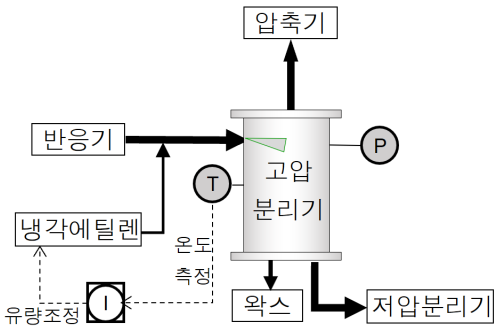

그림 4: 분리 공정

2.1.4 압출 및 제립 공정

압출기로 공급된 LDPE 수지는 분리기가 적정 레벨이 유지되는 속도로 압출된다. 압출기에 설치된 AMI 측정장치는 10분 간격으로 MI를 측정해 운전원이 물성 조절에 참조할 수 있도록 실시간으로 데이터를 전송한다. 압출기 말단 다이 플레이트(die plate)를 통과한 용융상태의 LDPE는 물을 공급해 굳힌 후 일정한 크기로 잘라 펠릿 형태로 제립 한다. 이때 다이 플레이트 전후 압력을 측정하며 이 또한 분자량을 가늠할 수 있는 운전 지표로 활용된다.

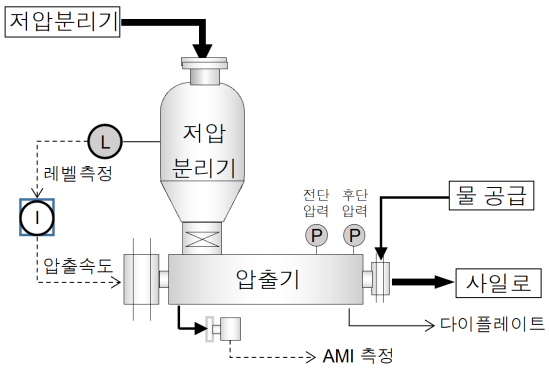

그림 5: 압출 공정

2.1.5 저장 및 포장 공정

펠릿은 탈수 후 건조과정을 거쳐 20톤 규모의 1차 사일로로 이송된다. 이송 중간에 위치한 샘플 테이블에서 실험실은 2시간 간격으로 제품을 샘플링하여 분석하며 분석 결과가 나오기 전까지 제품은 1차 사일로에 그대로 보관한다. 분석 결과 물성이 정상 범위로 확인되면 120톤 규모의 2차 사일로로 이송한다. 같은 2차 사일로에 보관한 제품은 동일 물성을 가

진 제품으로 간주하며 동일 lot를 부여한다. 고품질 제품 생산을 위해서는 매시간 제품 물성을 균일하게 생산하여 lot내, lot간 산포를 줄이는 것이 중요하다. 마지막으로 2차 사일로에 저장된 제품은 포장 공정으로 이송되며 출하 형태에 따라 25kg 또는 500kg 단위로 포장한 뒤 최종 출하된다.

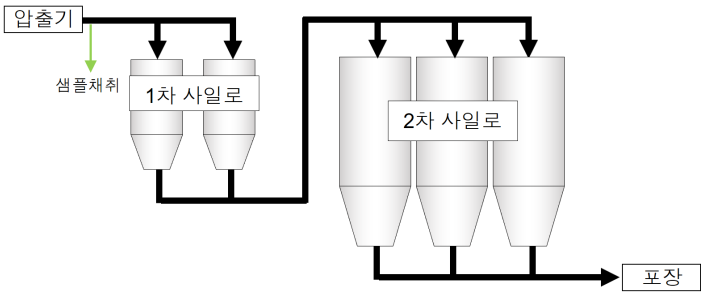

그림 6: 저장 및 포장 공정

2.2 LDPE 물성

LDPE 주요 물성은 분자량, 분자량 분포, 밀도 크게 세가지다.

첫번째, 분자량은 제품 종류를 분류하는 대표적 특성으로 상업적으로 MI를 이용해 측정하여 분자량을 구분한다. 일반적으로 분자량이 클수록 기계적 물성이 더 우수하며 가공성은 떨어지는 특징이 있다. 두번째, 분자량 분포는 반응기 종류 및 반응조건에 따라 결정되며 조정이 어렵다. 넓은 분자량 분포의 제품은 상대적으로 기계적 물성은 떨어지나 가공성이 우수하며 분자량 분포가 좁은 제품은 반대로 기계적 물성은 우수하나 가공성이 떨어진다. 마지막으로 밀도는 가장 조절하기 어려운 인자로 프로세스를 따라 결정된다. 결론적으로 제품 물성은 분자량 영향을 가장 크게 받으며 상업적으로 분자량 조절 기술이 가장 핵심적인 생산기술이다.

표 1: 분자량, 분자량분포, 밀도에 따른 일반적 물성 변화

구분	기계적물성	가공성(흐름성)
분자량↑	개선	감소
분자량분포↔	감소	개선
밀도↑	개선	감소

2.2.1 분자량

LDPE 수지는 에틸렌 단량체가 수만개 이상 모여 형성되며 분자량 분포는 정규분포를 띄고 있다. 이를 대표하기 위해 평균 분자량으로 분자량을 나타내며 상업적으로는 측정이 용이한 MI를 이용해 분자량을 표현한다. MI는 고분자 수지의 용융 점성도를 나타내는 지수이며 190°C 온도에서 용융된 LDPE를 2.16g의 추로 눌러 2.095mm의 오리피스를 통과한 양을 g/10min으로 환산하여 사용한다. 공정 압출기에 설치된 AMI 장치는

실시간으로 오리피스를 빠져나오는 수지량을 환산해 측정하나 압출 속도 등의 공정조건(외란) 영향을 많이 받으며 합부 판정 기준인 실험실 측정값과 차이가 있어 MI 측정이 없는 시간에 참조용으로만 사용한다. MI는 1이하부터 50까지 다양하며 숫자가 낮을수록 흐름성이 좋지 않은 고분자량 제품이다. 또한 MI가 작을수록 변동폭이 작으며 MI 7 전후의 제품에서 가장 민감하게 제품 물성이 움직인다.

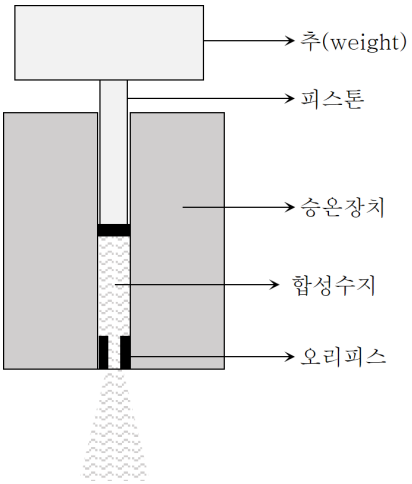

그림 7: MI 측정 장치 개요

2.2.2 분자량 분포

LDPE 제조공정은 반응기 형태에 따라 크게 autoclave 와 tubular 두 종류로 구분되며 특정 용도를 가진 제품의 분자량 분포는 그림 8과 같이 공정 라이선스가 결정되면 다른 라이선스에서는 기술적으로 생산이 불가능하다[2]. 상업적으로 단일 제품 내 분자량 분포는 동일한 것으로 간주한다. 임의로 고분자와 저분자 함량을 미세 조절하기 위해서는 반응기 온도 프로파일(profile) 조정 등이 필요하며 미세 조절만으로도 기계적물성, 가

공성 같은 물리적 성질이 크게 변한다.

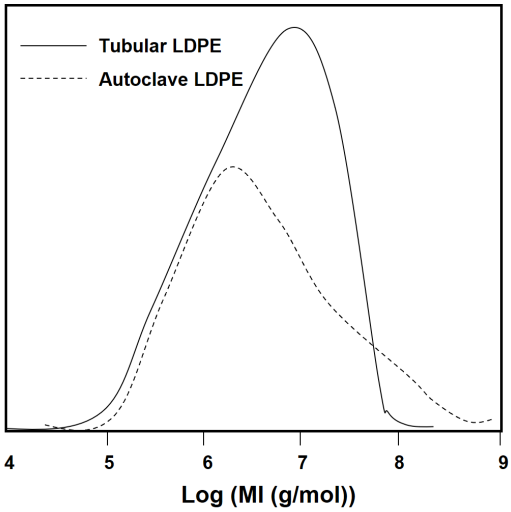

그림 8: 반응기 형태에 따른 분자량 분포 예시

2.2.3 밀도

LDPE는 0.91-0.93g/ml 범위 내의 밀도를 가지며 상업적으로 반응기 온도, 압력 조건이 정해지면 소수점 3자리 수준에서 결정된다. LDPE 제조 공정에서 해당 범위를 뛰어넘는 밀도 조절은 불가능하며 만약 다른 밀도의 합성수지가 필요하다면 필요 밀도를 고려하여 고밀도 폴리에틸렌(0.94-0.96)이나 선형저밀도 폴리에틸렌(0.91이하)과 같은 다른 프로세스에서 생산된 제품을 선택해야 한다.

2.3 기계 학습 알고리즘

기계 학습 중 지도학습은 독립변수와 종속변수의 관계를 학습을 통해 모델링하며 대표적으로 예측과 분류 모델이 있다. 예측 모델은 회귀분석을 기반으로 한 모델이 대표적이며 본 연구에서는 독립변수가 두 개 이상인 모델에 적합한 다중회귀분석을 사용하였다. 분류 모델은 기계 학습 및 딥러닝 분야에서 활발히 연구되는 분야로 간단한 분류기로 구분되는 나이브 베이즈, k-최근접 이웃, 의사결정나무부터 배깅과 부스팅, 랜덤포레스트, 서퍼트 벡터 머신, XGBoost 같은 좀 더 복잡한 학습이 가능한 모델이 있다. 시계열 분석은 종속변수만으로 과거값 또는 과거의 오차를 활용해 미래값을 예측하며 통계적 분석에 기반한 ARIMA 모델과 대표적 딥러닝 모델인 RNN에서 파생된 LSTM 알고리즘이 있다.

2.3.1 다중회귀분석

회귀분석은 연속형 변수들에 대해 두 변수 간의 관계를 수식으로 나타내는 분석 방법이며 독립변수가 2개 이상일 경우에 다중회귀분석을 사용하며 수식으로 표현하면 다음의 형태로 나타낼 수 있다.

$$y = a + bx_1 + cx_2 + dx_3 + \cdots + \varepsilon \tag{2.1}$$

2.3.2 나이브 베이즈

나이브 베이즈 분류는 조건부 확률을 구하는 베이즈 정리를 적용한 확률 분류 기법이다. 조건부 확률은 사건B가 일어났다는 조건하에 사건A가 일어날 확률을 P(A|B)로 표현하며 사후 확률이라고 한다.

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} = posterior = \frac{likelihood \times prior}{evidence} \tag{2.2}$$

여기서 P(A), P(B)는 각각 사건 A,B가 일어날 확률이며 베이즈 정리는 사건B가 발생하면 사건A가 발생할 확률이 어떻게 변하는지 표현한 식으로 B라는 사건을 관찰해 A가 발생할 확률이 어떻게 되는지 찾아내는 방법이다. 나이브 베이즈 알고리즘은 데이터가 클래스에 속할 확률을 베이즈 정리를 기반으로 계산한 후 클래스를 예측하게 된다[3].

2.3.3 k-최근접 이웃

k-최근접 이웃(k-nearest neighbor)은 가장 간단한 기계학습 알고리즘이며 새로운 데이터에 대해 이와 가장 거리가 가까운 k개의 과거 데이터의 결과를 이용해 다수결로 분류하는 방법이다.

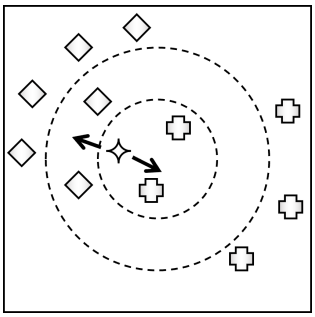

그림 9: k값 변화에 따른 데이터 분류

새로운 데이터에 더 가까운 이웃일수록 더 먼 이웃보다 평균에 더 많이 기여하도록 가중치(weight)을 줄 수 있으며 가중치 부여 방식에 따라 모델의 성능이 크게 달라질 수 있다[4].다양한 거리 계산 방법을 통해 가중치를 부여할 수 있으며 대표적으로 그림10과 같은 유클리디안, 민코프스키, 맨해튼 방법이 있다.

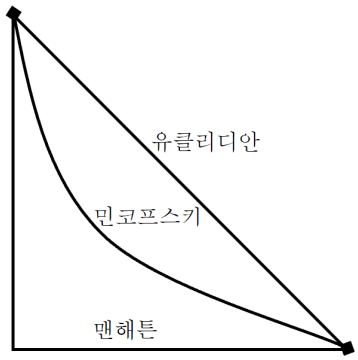

그림 10: 거리 계산 방법

$$\begin{aligned} \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} &\quad \text{유클리디안} \\ \sum_{i=1}^{k}\left(|x_i - y_i|^2\right)^{1/p} &\quad \text{민코프스키} \\ \sum_{i=1}^{k}|x_i - y_i|^2 &\quad \text{맨해튼} \end{aligned} \tag{2.3}$$

2.3.4 의사결정나무

주어진 독립변수에 의사결정규칙을 적용해 종속변수를 예측해 나가는 알고리즘이다. 분석 결과가 조건 형태로 출력되므로 모델을 이해하기 쉬운 장점이 있다. 그림11의 마름모 형태를 의사결정 노드라고 하며 특히 최상단에 위치한 것을 뿌리 노드라고 부른다. 의사결정 노드에서 true, false를 통해 종속변수는 다음 단계로 분류된다. 분류된 노드는 잎사귀 노드로 의사 결정이 종료되거나 다른 의사결정 노드로 이동해 다시 한번 분류가 이뤄지게 된다[5]. 의사결정나무는 의사결정 노드의 깊이 및 분류 조건에 따라 성능이 결정된다.

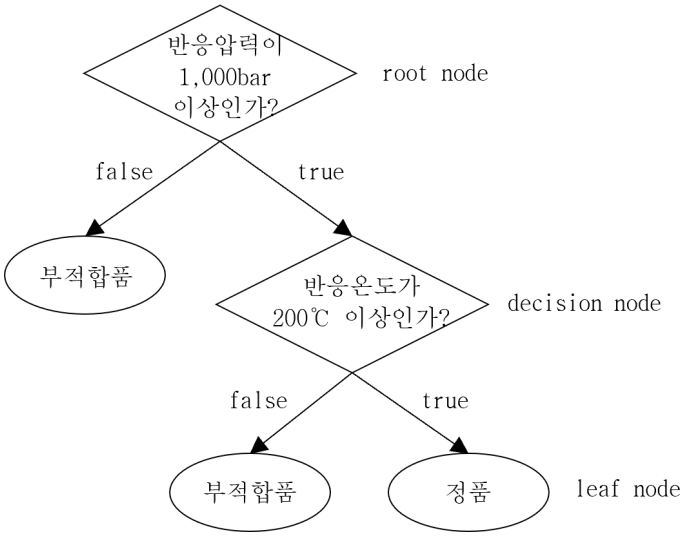

그림 11: 의사결정나무 예시

2.3.5 배깅& 부스팅

배깅과 부스팅은 각각 앙상블(ensemble) 모형 중 하나로 앙상블 모형은 여러 개의 예측 모델 조합해 최적화된 최종 예측 모델을 만든다.

배깅은 bootstrap aggregating에서 유래된 명칭으로 학습 데이터로부터 단순 랜덤 추출을 사용해 여러 개의 동일한 크기 샘플을 만들고(부트스트랩), 각 샘플 예측 모델 조합으로 최종 예측 모델을 만드는 방법이다.

부스팅은 배깅과 달리 각 샘플에 동일한 확률을 부여해 분류하지 않고 잘못 분류된 샘플에 더 큰 가중치를 적용해 새로운 분류 규칙을 만든다. 이 과정을 반복해 최종 모형을 만들게 된다 [6].

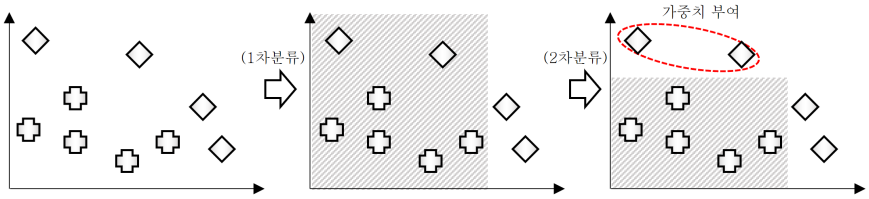

그림 12: 부스팅예시

2.3.6 랜덤포레스트

랜덤포레스트는 의사결정나무에 앙상블 기법인 배깅을 적용한 알고리즘이다. 단순 랜덤 추출을 이용해 동일 크기의 여러 트리 모델을 만들어 그 결과를 취합하고, 분류 모델은 다수결로 최종 모형을 출력한다. 배깅과의 차이점은 의사결정나무의 가지가 분할되는 노드 결정 방법에 있다. 배깅은 각각의 노드에서 모든 독립 변수를 이용해 최적의 분할 기준(노드)을 찾으나 랜덤포레스트는 각 노드마다 독립변수를 랜덤으로 추출한 뒤 추출된 독립변수만을 이용해 최적화를 실시한다[7]. 이러한 차이로 인해 일반적으로 배깅대비 랜덤포레스트의 예측 성능이 우수하다.

2.3.7 서포트 벡터 머신

Support vector machine(SVM)은 고차원인 n차원의 공간에서 자기보다 하나의 차원이 낮은(n-1) 최적의 분리 초평면을 찾고 초평면을 이용해 분류와 회귀를 수행하는 알고리즘 이다[8].

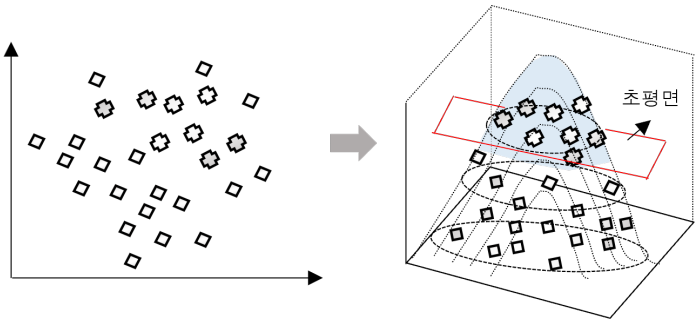

그림 13: 초평면예시

SVM은 서포트 벡터, 결정 경계, 마진등으로 구성된다. 서포트 벡터는 주어진 데이터 중에서 결정 경계와 가장 가까운 거리에 위치한 데이터를 말하며 결정 경계는 데이터의 분류 기준이 되는 경계이다. 그림13의 좌측 그림 같이 결정 경계를 찾기 어려운 경우는 커널 함수를 이용해 더 높은 차원의 데이터로 변형시켜 결정 경계를 찾을 수 있다.

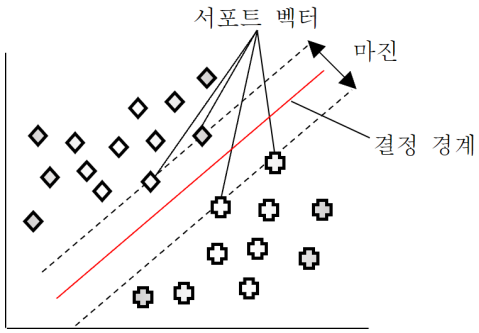

그림 14: 서포트 벡터 머신 구성 요소

2.3.8 XGBoost

XGBoost는 extream gradient boosting의 약자이다. XGBoost는 경사 하강법(greadient descent)을 이용해 순차적으로 틀린 것에 가중치를 부여해 모델의 성능을 개선하는 gradient boosting 알고리즘을 기반으로 한다. XGBoost 알고리즘은 함수의 기울기를 구하고, 기울기의 절대값을 낮은 쪽으로 계속 이동시켜 함수의 기울기가 0에 가깝게 될 때까지 반복하게 된다[9]. XGBoost는 다른 알고리즘 대비 조정 가능한 하이퍼 파라미터를 많이 갖고 있으며 이를 활용해 모델의 성능을 획기적으로 개선할 수 있다.

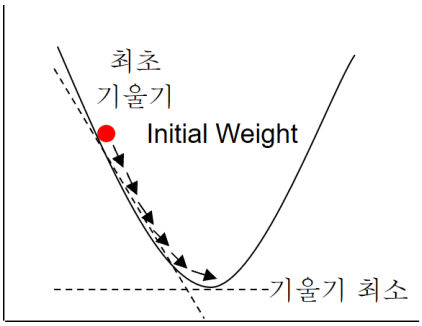

그림 15: 경사 하강법 개념

2.3.9 ARIMA

과거 시점의 값들을 독립변수로 하여 y를 예측하며, p시점 전까지 고려하는 AR(p) 자기 회귀(auto regression) 모형을 식(2.4)와 같이 정의한다.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{p-1} + \varepsilon_t \tag{2.4}$$

과거의 오차들을 독립변수로 사용하며 q시점 전까지의 오차항까지 고려하는 MA(q) 이동 평균(moving average)모형은 식(2.5)와 같다.

$$y_t = c + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \cdots + \phi_q \varepsilon_{1-q} + \varepsilon_t \tag{2.5}$$

AR 모형과 MA 모형이 혼재되어 있는 형태인 ARMA 모형은 시계열 자료가 안정적일때만 쓸 수 있으나 차분(differencing)을 구해 추세나 계절성을 제거 또는 감소시켜 식(2.6)의 ARIMA(auto regressive integrated moving average) 모형을 만들면 불안정한 시계열에도 적용할 수 있게 된다.

$$\begin{gathered} y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} \phi_1 \varepsilon_{t-1} + \phi_q \varepsilon_{q-1} + \varepsilon_t, \\ \textit{p=자귀회귀(AR)부분의 차수} \\ \textit{d=1차 차분이 포함된 정도} \\ \textit{q= 이동 평균(MA) 부분의 차수} \end{gathered} \tag{2.6}$$

식(2.6)을 후방이동(backshift) 기호를 이용해 식(2.7)과 같이 나타낼 수 있으며 이것을 ARIMA(p,d,q) 모델이라고 부른다[10].

$$\begin{gathered} B(By_t) = y_{t-1} \\ (1 - \phi_1 B - \ldots \phi_p B^p)(1-B)^d y_t = c + (1 + \theta_1 B - \cdots + \theta_q B^q)\varepsilon_t \end{gathered} \tag{2.7}$$

2.3.10 LSTM

LSTM은 recurrent neural network(RNN)의 기울기 소실을 해결하기 위해 고안된 시계열 분석 딥러닝 알고리즘이다. RNN은 과거 정보가 현재에 영향을 주는 시계열 분석에 주로 활용되고 있으며 LSTM은 RNN의 hidden state에 추가한 forget gate와 input gate가 메모리 역할을 수행해 RNN의 기울기 소실을 보완한다[11].

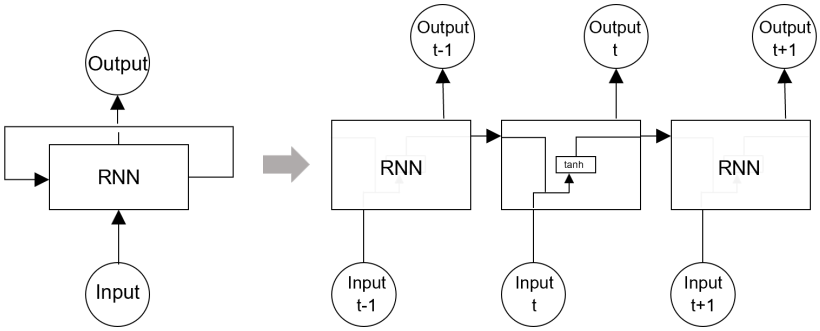

그림 16: RNN 구조

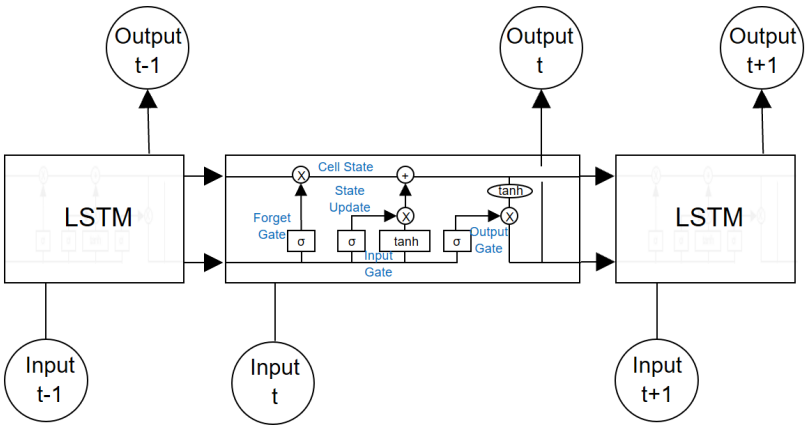

그림 17: LSTM 구조

제 3 장

연구 방법

3.1 연구 대상 선정

기계 학습 모델에서 예측할 종속변수는 MI이며 실험실 측정값을 기준으로 기계학습 모델 연구에 사용했다. 공정 중 용융지수 조절과 모니터링에 직접적인 관련이 있는 반응, 분리, 압출 및 제립 공정을 독립변수 연구 대상으로 한정하였다. 세 공정의 데이터 중 실제 운전 방법을 기반으로 다음과 같이 기계 학습 모델 구현을 위한 연구 대상을 최종 확정하였다.

표 2: 연구 대상 선정 결과

구분	연구 대상	관련 데이터
종속변수	MI	실험실 MI 측정값
독립변수	반응,분리,압출제립	공정 센서 및 AMI측정값

3.2 연구 방법

공정에 설치된 센서 종류는 크게 온도, 압력, 유량 센서가 있다. 연구 대상으로 선정한 세 공정에는 100여개의 센서가 설치되어 있으나 안전 및 단순 모니터링 목적으로 사용되는 센서를 제외하면 MI 조정에 직접 관련된 것은 20개 이하이다. 이 중 17개 센서 데이터를 운전 변수로 선정하여 6개월(21년 6월 - 21년 12월) 데이터를 수집하였다. 수집한 데이터 중 공정

및 센서 고장에 의한 이상치를 제거하고 실험실 물성 분석 간격인 2시간 기준으로 전처리를 실시하였다. 전처리 이후 물성데이터인 MI와 상관관계가 높은 공정 데이터 변수를 찾아 선정하고 7:3 비율로 트레인 및 테스트 세트로 분리하였다. 준비된 데이터 세트를 이용해 다음 세가지 방법으로 기계 학습 모델을 구축 후 공정 압출기에 설치된 실시간 AMI 장치와 기계 학습모델로 예측한 MI를 실험실에서 측정한 참값과 비교하였다.

첫번째, MI를 연속형 데이터 형태의 종속변수로 사용하고 공정 인자를 이용해 다중 회귀 분석 예측 모델을 생성하여 비교하였다.

두번째, MI를 정품 및 불량품 스펙 기준으로 이산형 데이터로 변형시켜 종속변수로 사용했으며 연속형 모델과 동일한 공정 변수를 이용해 나이브 베이즈, k-최근접 이웃, 의사결정나무, 배깅, 부스팅, 랜덤포레스트, 서포트벡터머신, XGBoost의 분류 계열 모델을 학습하고 최적화한 예측 성능을 AMI의 MI 예측값과 비교하였다.

세번째, MI 데이터를 이용해 시계열 모델인 ARIMA와 LSTM 모델을 학습해 MI를 예측해보고 모델 성능을 확인했다.

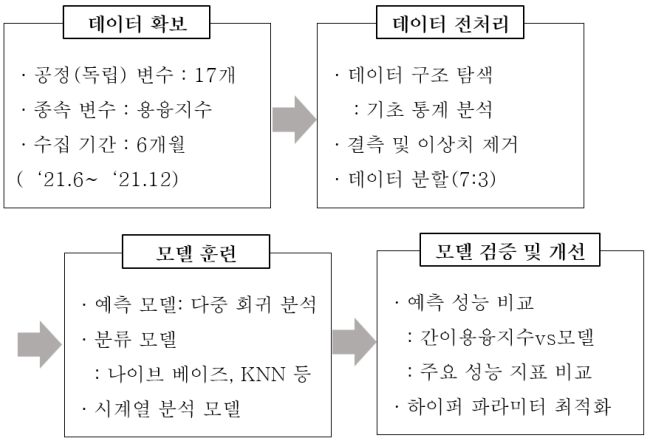

그림 18: 연구 방법 흐름도

3.3 연구 대상 평가 지표

예측하고자 하는 MI(종속변수)의 데이터 형태(연속형, 이산형)에 따라 평가 지표를 달리 적용하였다. 첫번째, MI를 연속형 데이터로 이용한 다중 회귀 분석의 경우 모델의 예측 수준을 파악할 수 있는 R^2와 오차 추정을 위한 RMSE를 평가지표로 선정했으며, 두번째, MI를 이산형 데이터로 이용한 나이브 베이즈등의 모델은 정확도와 함께 재현율을 모델 성능 판단 기준으로 삼았다.

3.3.1 결정계수 및 수정결정계수

결정계수 R^2는 모델의 설명력으로 해석하며 추정된 회귀식으로 실제 값을 얼마나 잘 설명할 수 있는지를 나타낸다. 0에서 1사이의 값을 가지며, 1에 가까울수록 예측력이 좋은 모델임을 의미한다.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad , (0 \leq R^2 \leq 1) \tag{3.1}$$

수정결정계수 R^2_{adj}는 독립변수의 수를 반영하여 수정된 결정계수로 서로 다른 수의 독립변수를 고려한 회귀 모형을 비교할 때 사용한다

$$R^2_{adj} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/[n-(k+1)]}{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{n-1}} \quad , k\text{=독립변수 수} \tag{3.2}$$

3.3.2 평균절대오차, 평균제곱근오차

평균절대오차 MAE와 평균제곱근오차 RMSE는 회귀모델의 예측값과 실제값의 차이를 이용해 예측 성능을 판단하는 지표이다.

MAE와 RMSE는 일반적으로 수치가 낮을수록 성능이 우수하며 예측 대상의 크기(scale)에 의존적이므로 다른 크기를 가진 모델을 서로 비교하기에는 적합하지 않다. 상대적으로 RMSE는 MAE 대비 특이값의 영향을 덜 받는 특징을 가지고 있다.

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.3}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.4}$$

3.3.3 혼동 행렬

기계 학습의 분류 문제에서 알고리즘의 성능을 시각화 할 수 있는 표이다. 혼동행렬을 통해 tp, fp, fn, tn 네 개의 결과를 얻을 수 있으며 이를 통해 정확도, 정밀도, 재현율을 구해 분류 모델의 성능을 파악할 수 있다[12]. 이진 분류에서 불균형한 클래스를 가진 데이터는 정밀도와 재현율을 상호 보완적으로 사용해야 한다. 일반적으로 모델 성능의 중요 지표로 제1종 오류를 고려하는 경우는 정밀도, 제2종 오류를 고려하는 경우는 재현율을 고려한다.

표 3: 분자량, 분자량분포, 밀도에 따른 일반적 물성 변화

		실제사실(actual)	
		참(p)	거짓(n)
예측값 (predicted)	참(p)	true positive(tp)	false positive(fp)
	거짓(n)	false negative(fn)	true negative(tn)

3.3.3.1 정확도

정확도는 positive와 negative를 정확하게 예측한 비율로 일반적인 분류 모델 성능 지표로 사용되며 직관적으로 이해하기 쉽다. 그러나 이진 분류에서 한 클래스로 편중된 불균형한 데이터의 경우는 평가 지표로 삼기 어렵다.

$$(accuracy) = \frac{tp + tn}{tp + fn + fp + tn} \tag{3.5}$$

3.3.3.2 정밀도

정밀도는 true로 예측한 것 중 실제 true인 값의 비율로 다음과 같다. fp(1종 오류)가 커질수록 정밀도가 낮아진다.

$$(precision) = \frac{tp}{tp + fp} \tag{3.6}$$

3.3.3.3 재현율

실제 true인 것 중 true로 예측한 값의 비율로 다음과 같다. fn(2종 오류)가 커질수록 재현율이 낮아진다.

$$(recall) = \frac{tp}{tp + fn} \tag{3.7}$$

3.3.3.4 AUC Score

AUC는 area under cover의 약자로 이진 분류 모델의 예측 성능을 판단하는 ROC(reciver operating charateristic) curve에 기반하여 AUC score를 계산하며 수치가 1에 가까울수록 분류 성능이 뛰어남을 의미한다.

ROC curve 그래프의 y축은 *recall*, x축은 flase positive rate(FPR)을 사용하며 FPR의 정의는 다음과 같다.

$$(FPR) = 1 - specificity = 1 - \frac{tn}{fp + tn} = \frac{fp}{fp + tn} \tag{3.8}$$

제 4 장

기계 학습 모델 구현 결과 및 분석

4.1 기계 학습 모델 구현 결과

MI와 상관관계가 높은 공정 변수를 17개 선정하였고 그 중 치명인자 12개를 재선별해 공정 변수로 이용한 모델과 성능을 비교해 보았다. MI는 데이터 형태에 따라 연속형, 범주형으로 나눠 기계학습 모델을 학습했으며 그 예측 결과를 비교하였다. 더불어 MI만을 독립변수로 사용하는 시계열분석 모델로도 MI를 예측해보았다. 모델로 예측한 MI값은 실제 운전원이 중요하게 참조하는 AMI 측정값을 공정 reference로 하여 AMI 로 예측한 MI 대비 개선율을 계산하였다.

연속형 데이터를 이용한 예측 모델의 경우 공정 reference는 R^2_{adj} 0.341이었으나 기계학습 모델은 0.455-0.457로 최대 34% 개선됨을 확인하였다. 범주형 데이터를 활용한 분류 모델의 reference는 정확도 0.953, 재현율 0.270이며 기계학습 모델은 정확도 0.916-0.969, 재현율은 0.211-0.556으로 모델별 다양한 결과를 나타냈다. 이 중 XGBoost에서 정확도 0.969, 재현율 0.556로 가장 높은 성능을 보였다. 시계열 모델은 LSTM의 유용성을 확인했으나 데이터 부족으로 유의한 MI 예측 결과를 얻지는 못했다.

4.2 데이터 전처리

LDPE 제조 과정은 연속 공정이며 공정데이터는 1분 간격으로 수집된다. 수집된 데이터는 이상반응, 설비 고장 등의 요인으로 이상 데이터가 필연적으로 발생하며 공정 전문가의 판단 하에 직접 제거해야 한다. 기초 통계 분석으로 데이터를 시각화하고 결측치 및 이상치를 확인한 후 데이터 구간의 상위 및 하위 5%를 벗어나는 이상 데이터를 제거하였다.

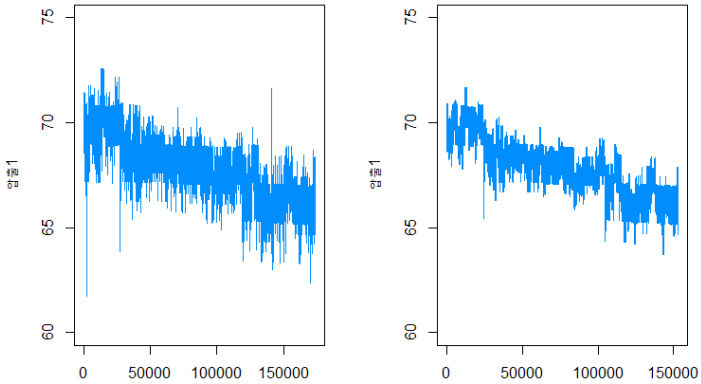

그림 19: 공정 이상데이터 전처리 전후

물성 데이터인 MI는 실험실에서 2시간 간격으로 측정되며 공정, 물성데이터의 서로 다른 수집 간격을 일치시키기 위해 MI 측정시간을 기준으로 공정 데이터의 구간 평균을 구했다. 표4와 같이 MI 측정 시점에 수집된 전처리 원본 데이터(raw) 값과 구간 평균값의 회귀식 설명력을 비교한 결과 구간 평균값의 설명력이 약1.5배 높음을 확인할 수 있었다.

표 4: 원본 vs 구간 평균 설명력 비교

데이터수(개/구간)	설명력(R^2_{adj})
원본(Raw Data)	30.6%
구간 평균	45.5%

4.3 예측 모델

앞선 4.2절에서 운전원이 실제 물성 조절시 참조하는 공정변수 17개를 선별하여 전처리 하였다. 선별된 변수는 반응공정 관련 변수 6개, 분리공정 관련 변수 6개, 압출 및 제립 공정 관련 변수 5개로 구성됐다. 반응 및 분리 공정 변수는 온도 및 압력을 조절하고 모니터링 하는 센서 측정값이며, 압출 및 제립 공정은 AMI 측정값을 제외하고 압출기 상태를 확인할 수 있는 센서 측정값이다.

선별한 공정변수와 MI 상관관계 분석 결과 AMI 측정값은 0.58의 강한 상관관계를 보였으며 그 외 인자는 유의한 상관관계를 보이지 않았다. 그러나 인자의 유의성 검증을 위한 회귀 분석에서는 12개의 인자들이 유

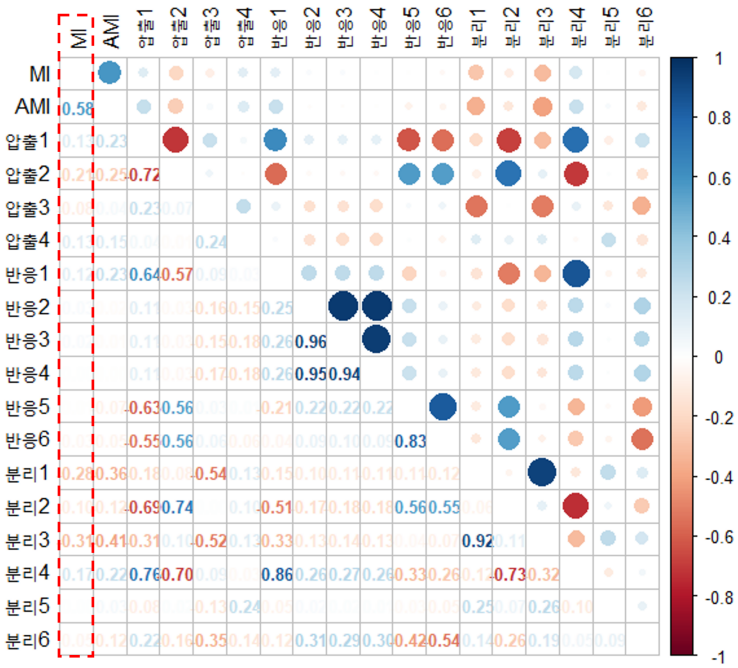

그림 20: 상관관계 분석 결과

의한 것으로 확인되었다. 상관계수 값의 판단 기준은 실무적 해석에 따라 달라질 수 있으므로 회귀분석 결과 유의한 인자는 모두 독립변수로 사용하였다. 다중회귀식의 종속변수는 MI, 독립변수는 앞서 분석한 상관 회귀분석을 바탕으로 12개를 선택하였고 성능 비교를 위해 17개의 독립변수를 이용한 모델도 구현하였다.

모델 성능 비교를 위해 AMI를 reference로 하여 예측한 모델의 R^2_{adj} 0.341, RMSE 0.086을 기준으로 삼았다. 독립변수 12개를 사용한 모델은 R^2_{adj} 0.457, RMSE 0.076으로 R^2_{adj} 34%, RMSE 12% 개선을 확인했으며 독립변수 17개를 사용한 모델에서도 유사한 결과를 확인할 수 있었다. 다중회귀분석 모델의 성능은 R^2_{adj} 0.457로 설명력이 너무 낮아 기존 공정의 AMI를 대체할 수는 없다. 다만 독립변수에 개수에 따른 연속형 데이터 모델 성능 차이가 없는 것을 확인할 수 있었고 이에 따라 운전 경험을 바탕으로 사전 선정된 17개의 공정변수를 분류 모델에서도 사용하기로 했다.

표 5: 독립변수별 모델 성능 비교

구분	R^2_{adj}(개선율)	RMSE(개선율)	MAE(개선율)
reference	0.341(-)	0.086(-)	0.068(-)
독립변수 12개	0.457(34%)	0.076(12%)	0.059(12%)
독립변수 17개	0.455(34%)	0.076(12%)	0.059(12%)

4.4 분류 모델

MI 스펙 기준으로 범주형 데이터로 변환 후 분류 모델을 구현하였다. 앞서 회귀식과 동일 독립변수로 로지스틱 회귀 분석 실시 결과 92%의 정확도를 보여 범주형 분류 모델의 가능성을 확인할 수 있었다.

전처리한 데이터를 7:3 비율로 트레인 및 테스트 데이터로 분할했으며 reference는 연속형 데이터와 동일한 AMI로 예측한 MI 값을 기준으로 했다. 단, 전체 MI 값 중 불량품 비중이 1.4% 수준으로 매우 낮은 불균형한 레이블 클래스를 가진 데이터이므로 재현율을 정확도와 함께 성능 지표로 선정하였다. 나이브 베이즈를 비롯한 총 8개의 기계 학습 모델을 구현 한 결과 모델별로 재현율에 큰 차이를 보였다(-22% +106%). 범주형 모델은 정도와 재현율의 최적값을 찾으며 스펙을 벗어난 데이터 수가 현저히 적은 경우 정밀도가 높을수록 재현율은 트레이드 오프(trade off)로 급격히 낮아져 모델별 재현율 편찬가 커진 것으로 판단된다. 분류 모델중 XGBoost에서 정확도, 정밀도, 재현율, AUC Score 모두 최대 성능을 보였으며 산업 현장에서 사용 가능한 수준이다.

표 6: 분류 모델별 성능 비교

구분	정확도	정밀도	재현율	AUC score
reference	0.953	0.571	0.270	0.630
나이브 베이즈	0.916	0.263	0.526	0.730
k-최근접 이웃	0.959	0.571	0.211	0.602
의사결정나무	0.948	0.357	0.263	0.621
배깅	0.957	0.500	0.316	0.651
부스팅	0.961	0.600	0.316	0.653
랜덤포레스트	0.968	1.000	0.263	0.632
서포트벡터머신	0.964	0.800	0.211	0.604
XGBoost	0.969	0.833	0.556	0.774

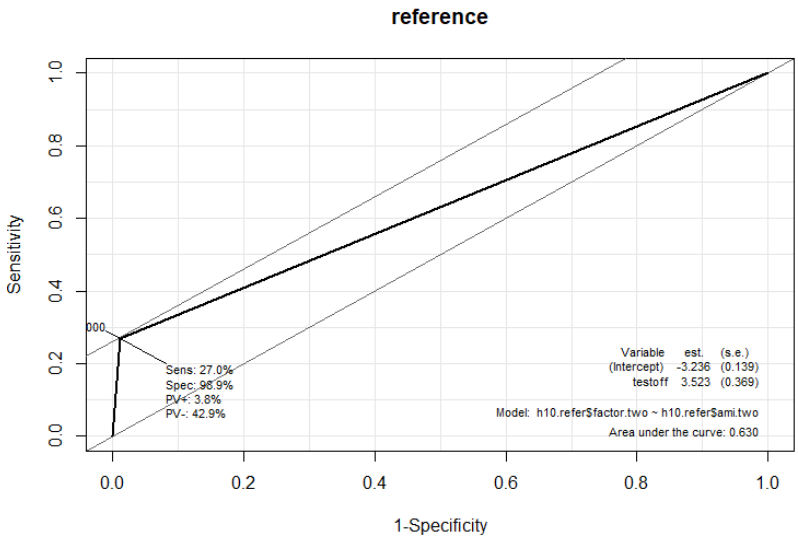

그림 21: Reference의 AUC

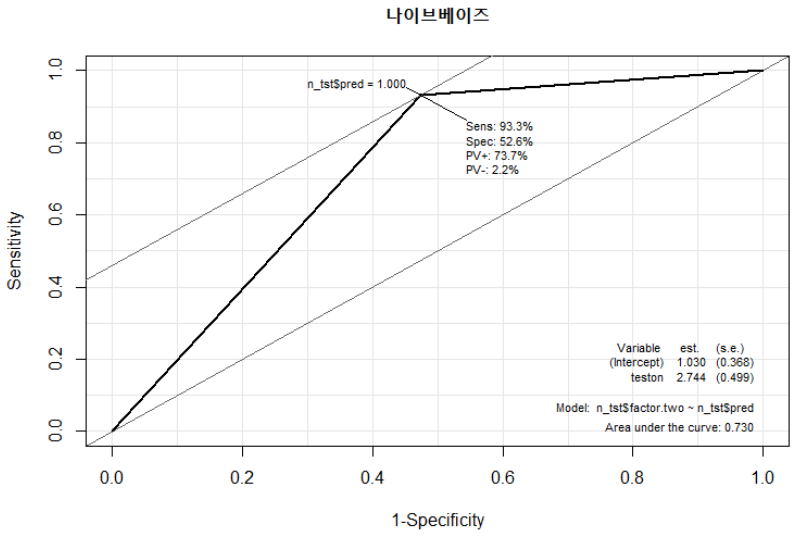

그림 22: 나이브베이즈의 AUC

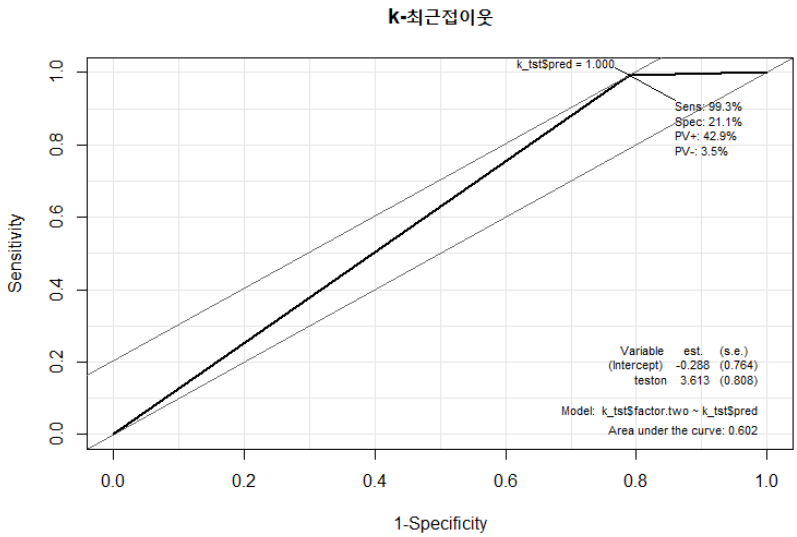

그림 23: k-최근접이웃의 AUC

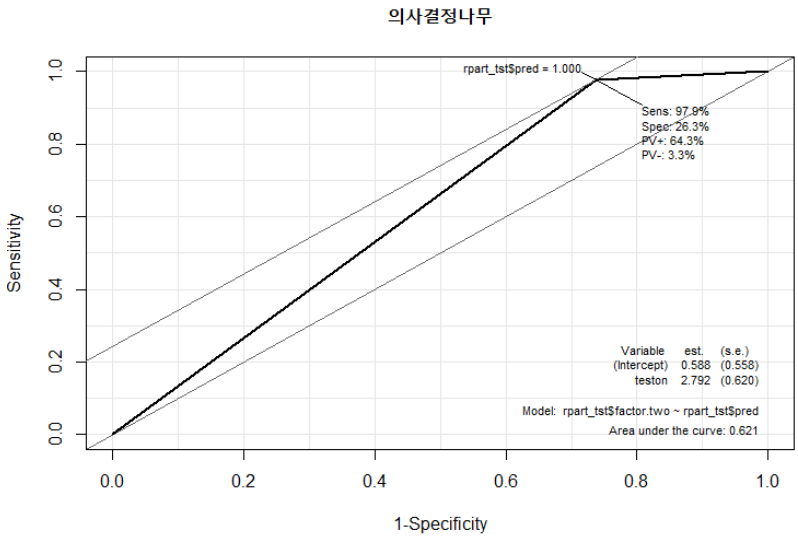

그림 24: 의사결정나무의 AUC

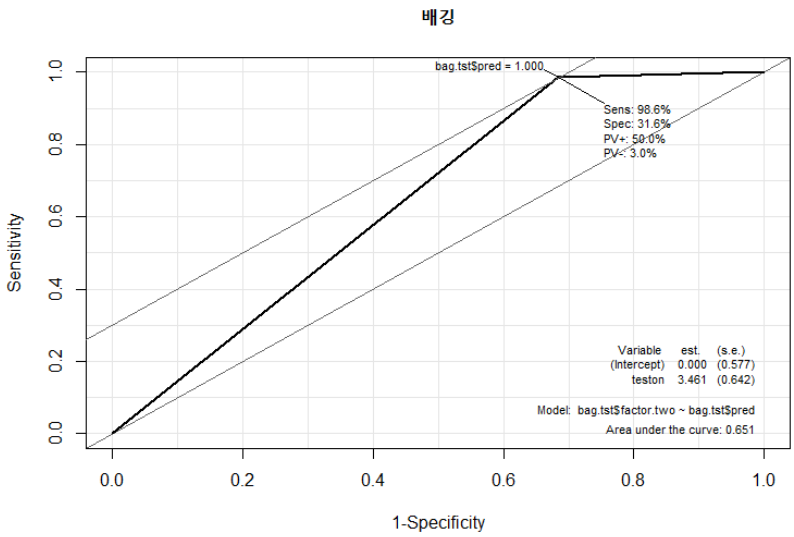

그림 25: 배깅의 AUC

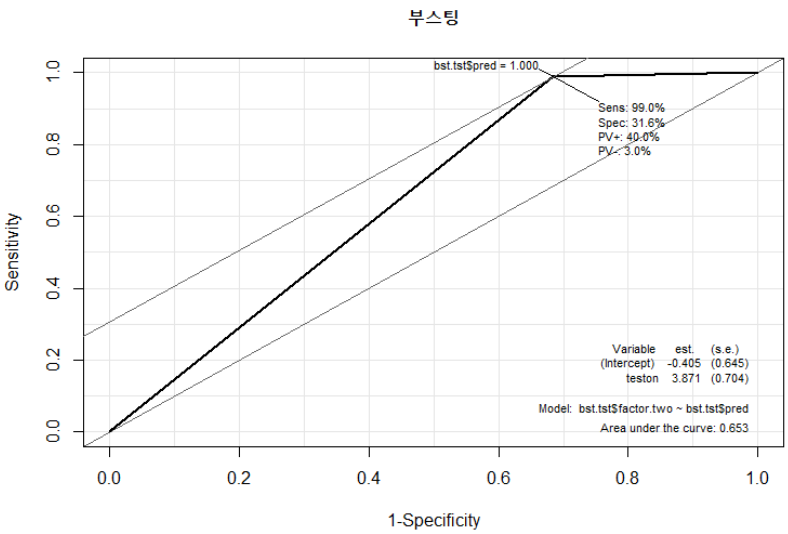

그림 26: 부스팅의 AUC

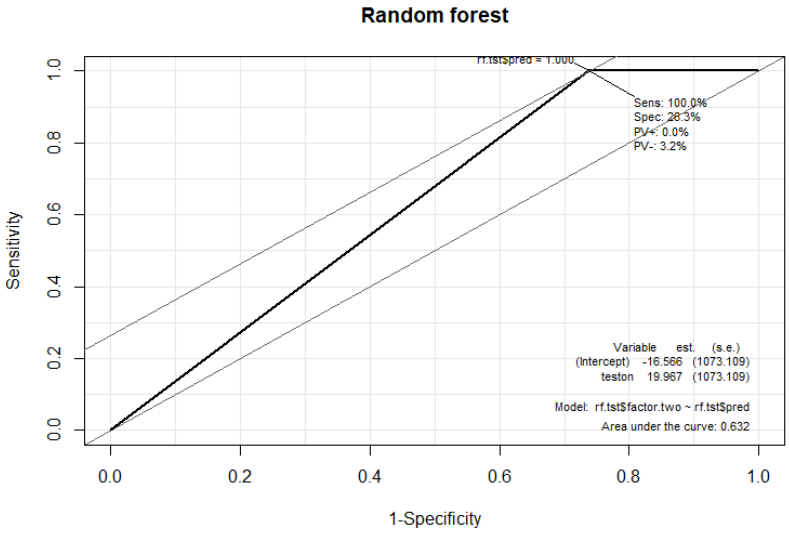

그림 27: 랜덤포레스트의 AUC

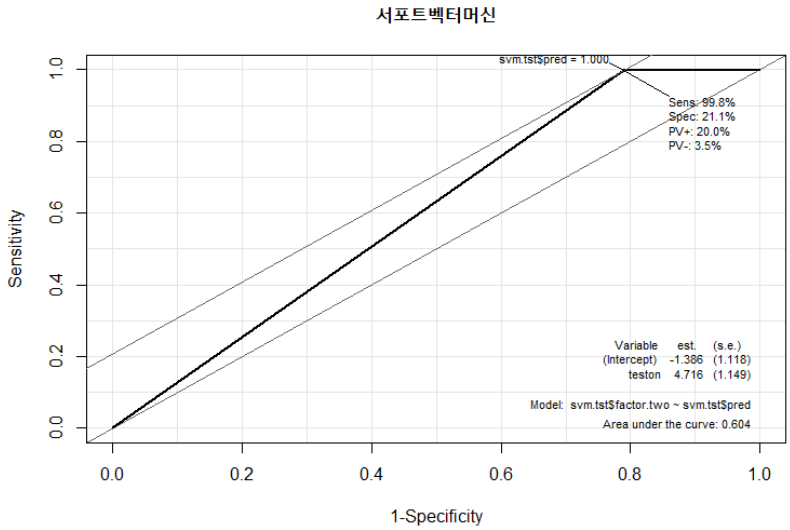

그림 28: 서포트벡터머신의 AUC

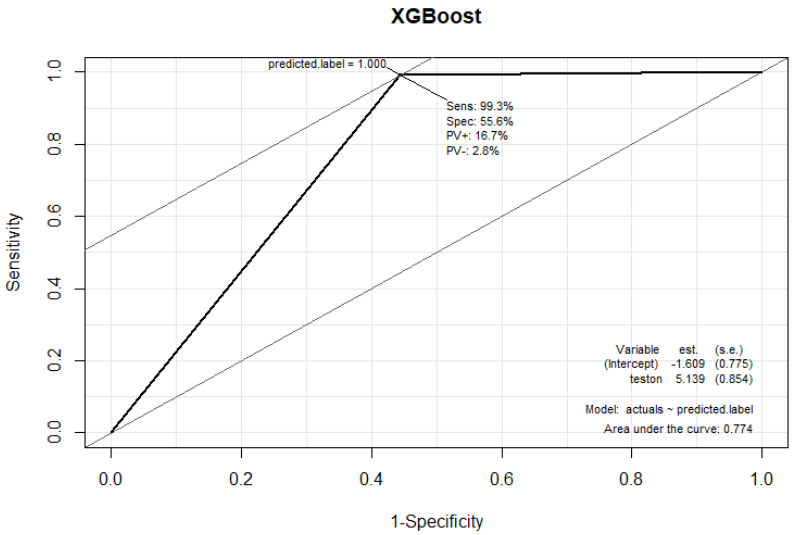

그림 29: XGBoost의 AUC

4.4.1 하이퍼 파라미터 최적화

모델은 학습에 사용되는 하이퍼 파라미터 최적화로 성능을 개선한다. 대부분 모델의 하이퍼 파라미터는 2-3 종류를 최적화하나 XGBoost는 상대적으로 다양한 하이퍼 파라미터를 갖고 있고 조정이 용이한 특징이 있다[13]. XGBoost의 주요 하이퍼 파라미터의 베이지안 최적화[14]를 통해 표7과 같이 본 연구 모델에 적합한 최적값을 도출했다.

표 7: XGBoost 하이퍼 파라미터

파라미터	별칭	범위	기본값	최적값
eta	learning rate	[0,1]	0.3	0.4
gamma	min split loss	[0,∞]	0	0
max depth	max depth	[0,∞]	6	10
min child weight	min child weight	[0,∞]	1	0
sub sample	sub sample	[0,1]	1	0.7
lambda	reg lambda	-	1	10
alpha	reg alpha	-	0	3.2

그림30의 cross validation을 통해 XGBoost의 학습 과정을 분석한 결과 iteration 134회에서 성능이 최적화되었을 알 수 있다.

그림31의 변수별 중요도에서는 AMI값이 예측에 가장 높은 중요도로 사용됨을 알 수 있다[15]. 두번째 중요 변수로 사용된 반응압력은 용융지수 결정에 가장 결정적인 영향을 주는 반응 인자이며 그 외에 용융지수 예측과 관련된 대부분의 압출 관련 변수들이 상대적으로 모델 학습에 중요인자로 사용되었다. 이는 운전원이 AMI값을 기준으로 공정 변수들을 조합하여 MI를 예측하고 조절하는 실제 용융지수 예측 과정과 동일하며 기계 학습 모델도 유사한 과정을 통해 MI값을 모델링함을 알 수 있다. 반면 반응 온도는 상대적으로 모델에 끼치는 영향이 작은데 이는 반응온도의 표준편차가 매우 작아(0.20) 분별력이 작기 때문으로 보인다. 온도

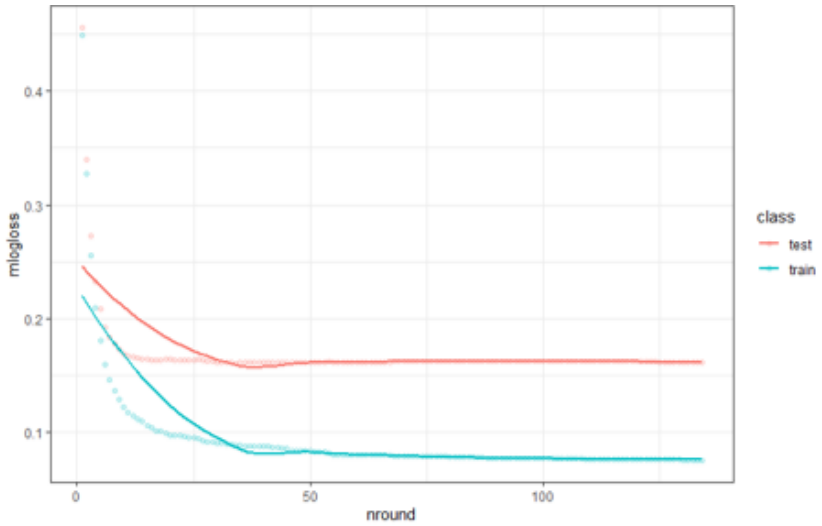

그림 30: XGBoost 모델 학습 결과

중에서도 RV_btm, RV_temp는 반응기 내부 온도가 아닌 반응기 하부와 출구온도로 실제 공정에서 물성 예측 활용 변수로 중요도가 가장 낮다. 모델에서도 동일하게 이 두 변수의 gain값이 중요도를 낮게 차지하는 점도 주목할 결과이며 실제 변수 중요도와 동일하게 머신 러닝 모델에 변수 중요도가 반영된 점이 모델을 더욱 신뢰할 수 있게 한다.

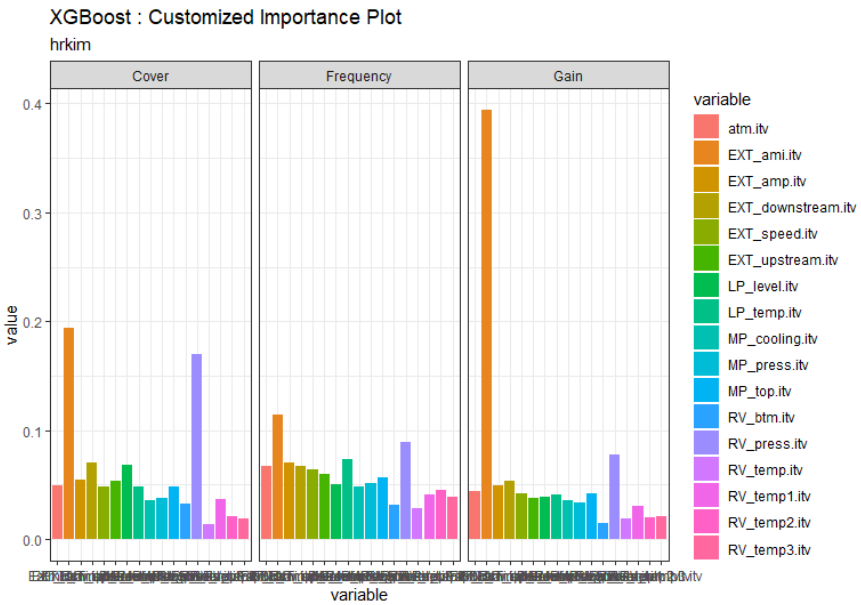

그림 31: XGBoost 변수별 중요도

4.5 시계열 모델

앞선 회귀 및 분류 모델은 독립변수와 종속변수와의 관계를 파악하는데 중점을 두고 있다. 하지만 연속 공정에서 생성된 LDPE의 MI 현재값 t는 직전값 t-1, t-2, … t-n과 일정한 경향성을 갖고 있으며 이는 과거 값을 이용해 현재 값을 예측하는 시계열 분석 방법을 사용할 수 있음을 의미한다. 시계열 모델 중 통계 기반의 ARIMA 모델과 딥러닝 기반의 LSTM 두 가지 모델로 MI를 예측하여 시계열 모델 적용 가능성을 확인하였다.

첫번째 ARIMA 모델은 ARIMA(0,1,1)의 MI 예측모델을 구현하였으나 예측 신뢰 구간이 넓어 실제 운전에 사용하기에는 무리가 있었다. 두번째 LSTM 모델의 MI 예측은 데이터 수 부족으로 유의한 결과를 확인하지 못했으나 MI와 유사한 데이터 산포를 가진 AMI 데이터로 예측 모델을 만들 수 있었으며 이를 바탕으로 현업 적용 가능성을 확인할 수 있었다.

4.5.1 ARIMA 모델

그림32의 방법론으로 ARIMA 모델 예측을 수행했다. 첫번째 데이터 준비 단계에서 그림33과 같이 수집된 MI 데이터로 분포 그래프 그려 분산 형태를 확인했으나 그래프에서는 비정상성 여부를 뚜렷하게 판단하기 불가했다. 따라서 동일 MI데이터로 Phillips-Perron unit root test[16] 정상성 검증을 실시해 비정상성 여부를 판단한결과 p-value가 0.05 이하로 유의하여 원(raw)데이터 자체가 안정화되어 있는 데이터임을 알 수 있었다.

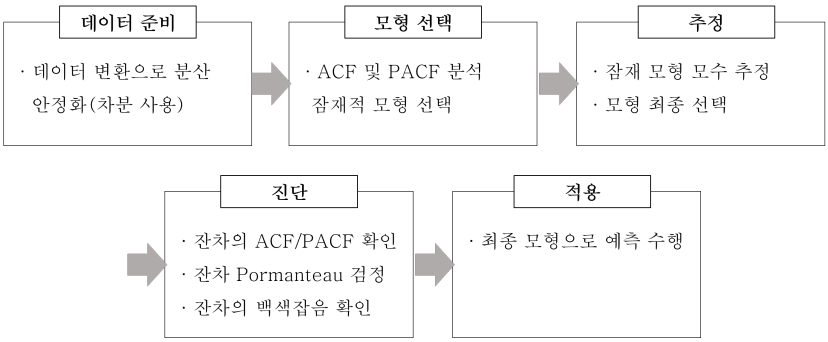

그림 32: ARIMA 모형 예측 방법

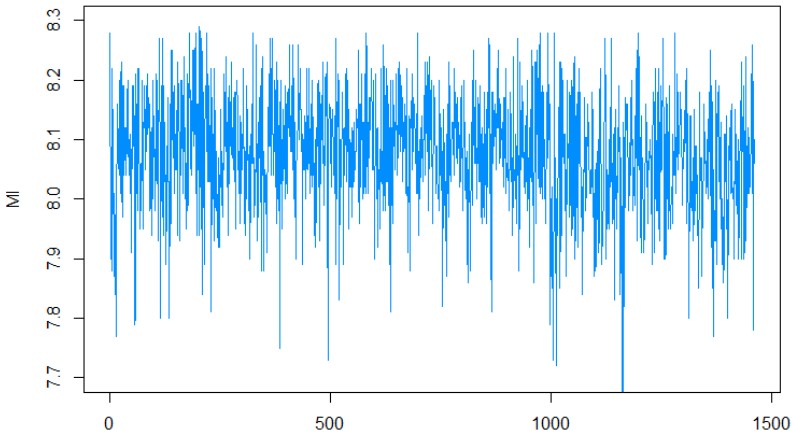

그림 33: MI 데이터 분포

```
        Phillips-Perron Unit Root Test

data:  ldpe
Dickey-Fuller Z(alpha) = -1304.7, Truncation lag parameter = 7, p-value = 0.01
alternative hypothesis: stationary
```

그림 34: 정상성 Test 결과

두번째 단계인 모형 선택을 위해 원데이터를 이용해 그림35와 같이 autocorrelation function(ACF)와 partial autocorrelation function(PACF) 그래프를 그려보았다. ACF 그래프 분석 결과 절단점이 5 이상이며 PACF에서도 그래프의 안정화된 모습을 확인할 수 없었다. 결과적으로 현재 데이터에서는 적합한 모형을 선택할 수 없어 데이터를 1차 차분하여 그래프 분석을 다시 실시하였다.

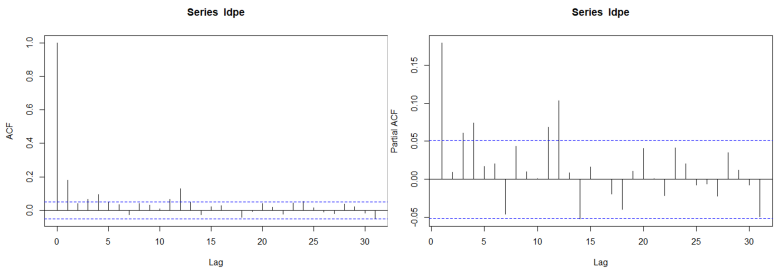

그림 35: Raw Data의 ACF, PACF

그림36의 1차 차분 데이터 그래프 분석 결과 원데이터와 달리 ACF 그래프에서 첫번째 스파이크 이후 값들이 0으로 빠르게 수렴함을 알 수 있으며 PACF 그래프도 0으로 수렴하는 형태를 확인할 수 있었다. 이를 통해 1차 차분 데이터의 MA(1) 모델을 잠재적 모델로 선택할 수 있었다.

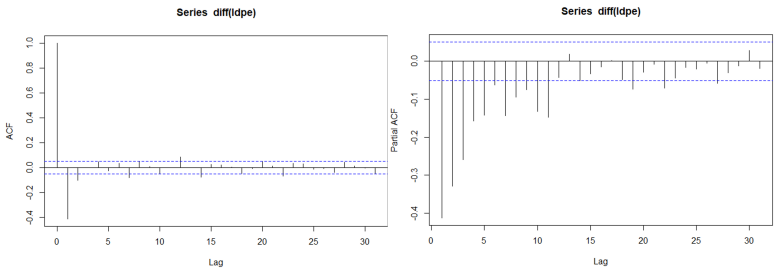

그림 36: 1차 차분의 ACF 및 PACF

세번째로 앞서 선택된 ARIMA(0,1,1) 잠재 모형의 적정성을 판단하기 위해 모수 추정을 실시했다. 추정 결과 likelihood, Akaike information criterion(AIC)는 그림37과 같았으며 잔차 진단 결과도 특별한 이상이 확인되지 않아 ARIMA(0,1,1) 모델을 최종 모델로 확정할 수 있었다.

```
summary(arima.test2_1)
Series: ldpe
ARIMA(0,1,1) with drift

Coefficients:
        ma1  drift
    -0.9878      0
s.e.  0.0074      0

sigma^2 = 0.009588:  log likelihood = 1320.88
AIC=-2635.77   AICc=-2635.75   BIC=-2619.91

Training set error measures:
                    ME      RMSE      MAE        MPE     MAPE     MASE     ACF1
Training set 0.0006614424 0.09781789 0.0770327 -0.006453842 0.9556297 0.7726539 0.1583568
```

그림 37: 모수 추정 결과

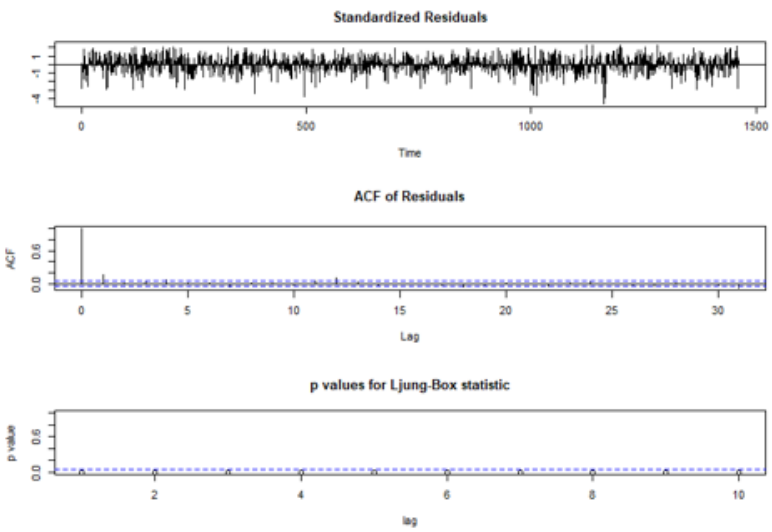

그림 38: 잔차 분석 결과

마지막으로 결정된 모델을 이용해 용융지수 예측을 실시했다. +5개의 값을 예측한 결과는 그림39와 같다. 1차 차분을 이용했으므로 1461번째 실측값 대비 차이값을 예측하게 된다.

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
1462	-7.668012e-03	-0.1331552	0.1178192	-0.1995841	0.1842481
1463	-1.578652e-05	-0.1764055	0.1763739	-0.2697805	0.2697490
1464	-1.578652e-05	-0.1764055	0.1763739	-0.2697805	0.2697490
1465	-1.578652e-05	-0.1764055	0.1763739	-0.2697805	0.2697490
1466	-1.578652e-05	-0.1764055	0.1763739	-0.2697805	0.2697490

그림 39: 예측 결과(수치)

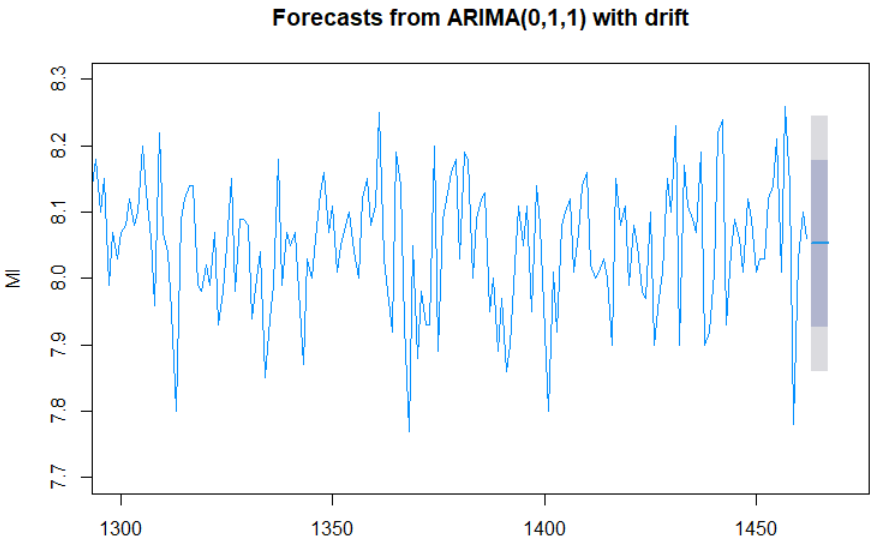

그림 40: 예측 결과(그래프)

최종 구현된 ARIMA 모델의 MI 예측값의 신뢰구간 95%를 고려할 경우 -0.20 +0.18 범위를 갖으며 이는 MI 공정 관리 범위(±0.2)와 유사한 수준이다. 현재 공정 MI의 분산이 매우 작기 때문에 ARIMA 모델이 넓은 신뢰 구간을 갖고 있는 것으로 판단된다.

4.5.2 LSTM 모델

LSTM 모델 분석에 앞서 원활한 모델링이 되도록 데이터 형태를 정규화 하였다. RNN 모델의 input layer shape은 12로 지정하여 2시간 간격으로 측정되는 MI의 24시간 데이터를 사용할 수 있게 하였고 hidden layer로는 LSTM을 정의하여 학습을 하였다[17].

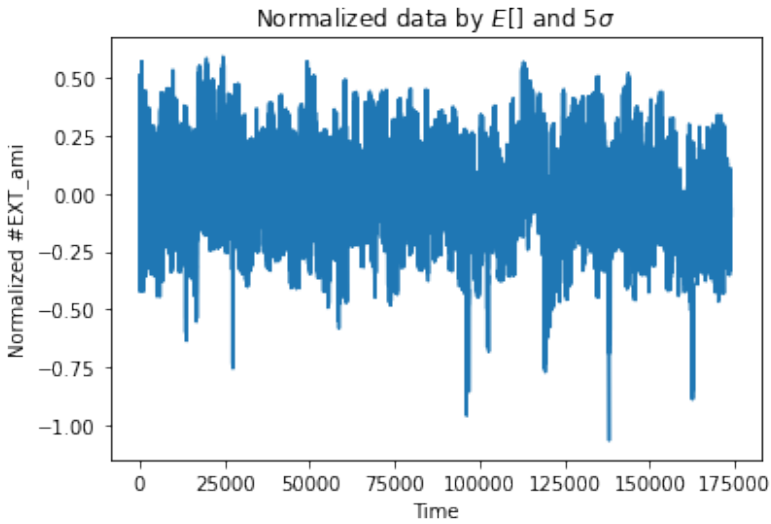

그림 41: 정규화된 그래프

그림42의 학습 결과 training loss 대비 validation loss가 줄어 들지 않아 학습이 원활하게 진행되지 않았음을 알 수 있었고 그림43과 같이 참값(original)과 예측값(prediction)의 편차 또한 큼을 알 수 있었다.

학습이 잘 되지 않은 원인은 크게 두 가지다. 첫번째는 데이터 자체 결함이며 두번째는 하이퍼 파라미터 최적화를 비롯한 모델 구현의 문제다. 두 가지 원인 중 근본적이며 가능성이 더 높은 데이터 자체 결함 확인을 실시 했다. 딥러닝은 그 특징상 학습에 다량의 데이터가 필요하나 앞서 구현된 모델에 사용된 MI 데이터는 상대적으로 적은 1462개 데이터가 사

그림 42: MI의 LSTM 모델 학습 결과

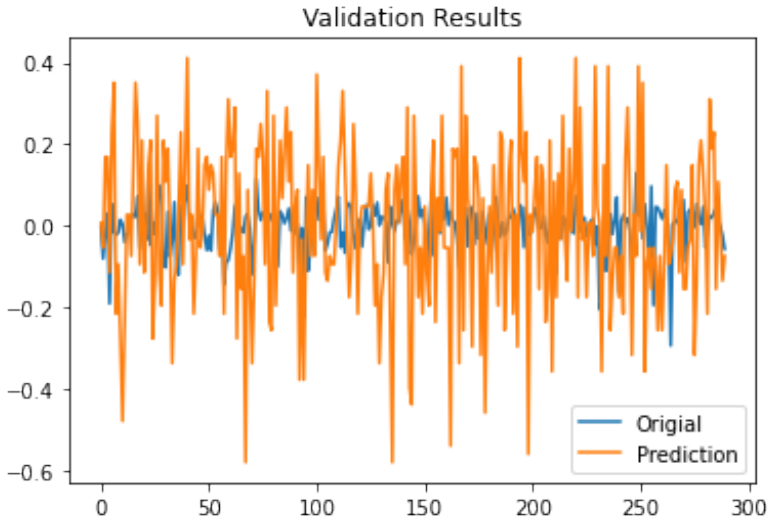

그림 43: MI의 LSTM 참값 vs 예측값

용되어 학습이 잘 진행되지 않았을 가능성이 높다. 따라서 데이터를 충분히 확보 후 재검증해야 하나 현실적으로 불가능하여 MI와 유사한 평균과

분산을 가진 AMI 데이터로 동일 모델 학습을 실시해 데이터 수가 충분히 확보되었을 때의 LSTM 모델 유효성을 판단했다. AMI는 공정 데이터로 2초 간격의 데이터 수집이 가능하며 동일 데이터 수집 기간에 MI대비 120배인 17,000개 이상의 데이터를 확보할 수 있다.

동일 모델로 학습 데이터만 AMI 데이터로 변경하여 학습을 진행한 결과 앞선 학습 결과와 달리 그림44와 같이 validation loss가 training가 유사한 수준으로 감소하여 최소값에 수렴하는 경향을 확인할 수 있었다. 그림45의 모델 예측값과 실제값을 비교한 그래프도 MI 모델 대비 정확도가 매우 높음을 확인할 수 있었다. 이러한 결과를 바탕으로 모델 자체 결함이 아닌 데이터 결함으로 모델 학습이 불량한것으로 판단되며 충분한 수의 MI 데이터가 확보된다면 LSTM 모델을 이용해 높은 정확도의 예측 모델을 구현할 수 있을 것으로 예상된다.

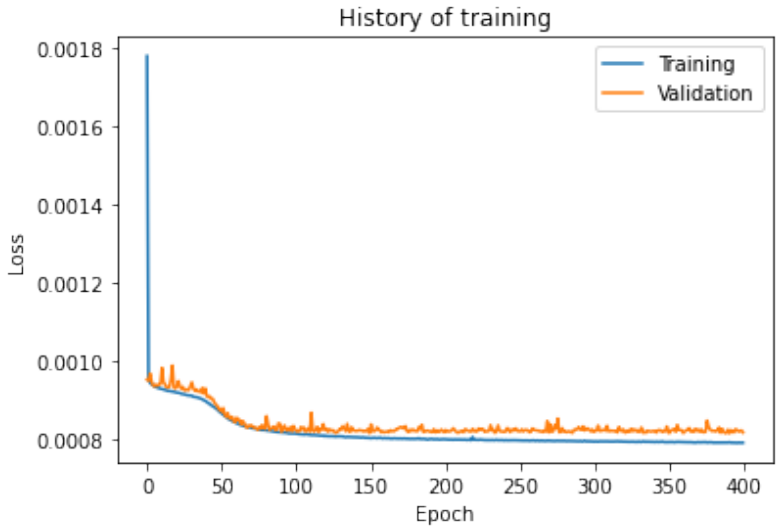

그림 44: AMI의 LSTM 모델 학습 결과

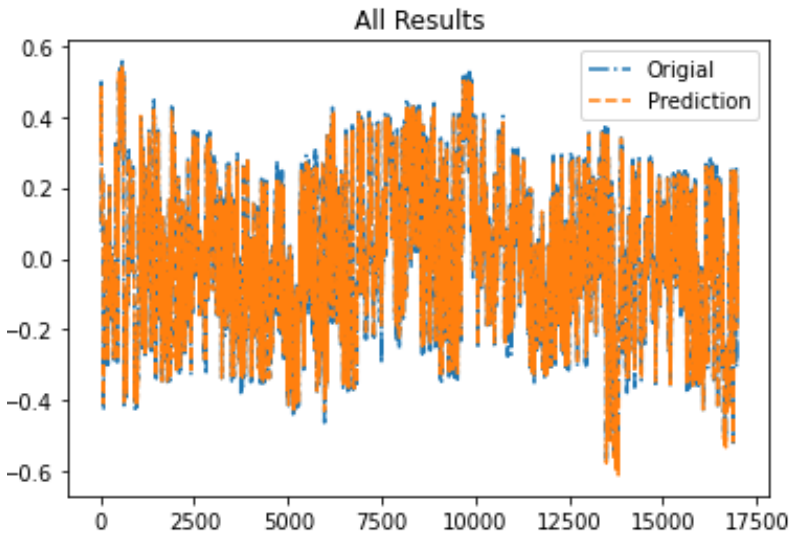

그림 45: AMI의 LSTM 예측값 vs 실측값

LSTM 모델을 이용한 미래값 예측은 단일값 예측과 여러개의 미래값을 예측하는 다중값 예측이 모두 가능하다. 그러나 단일 예측과 달리 LSTM 다중 예측 모델[18]은 예측 정확도가 떨어지며 데이터수가 증가할수록 예측에 필요한 하드웨어 자원이 기하급수적으로 증가하게 된다. 빠른 판단이 필요한 실제 생산현장에서는 예측에 필요한 자원과 시간을 고려해야 하며 다중 예측은 현실적으로 활용이 불가능하다. 현재 시간 t 보다 앞선 t+1 시간의 용융지수 예측만으로 현장 활용은 충분하므로 단일 예측 모델 중심으로 LSTM 추가 연구가 필요할 것으로 판단된다.

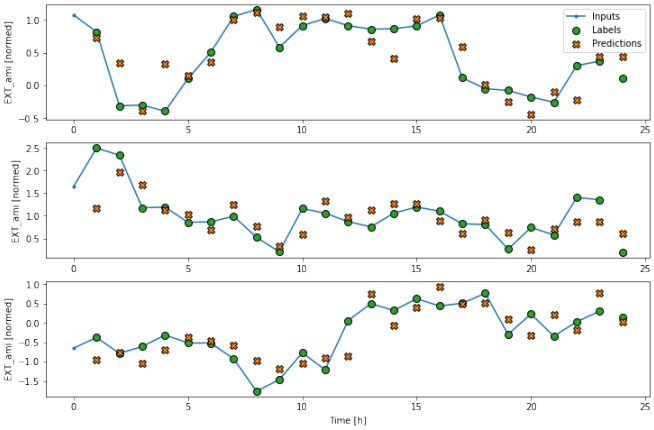

그림 46: LSTM 단일 예측 결과

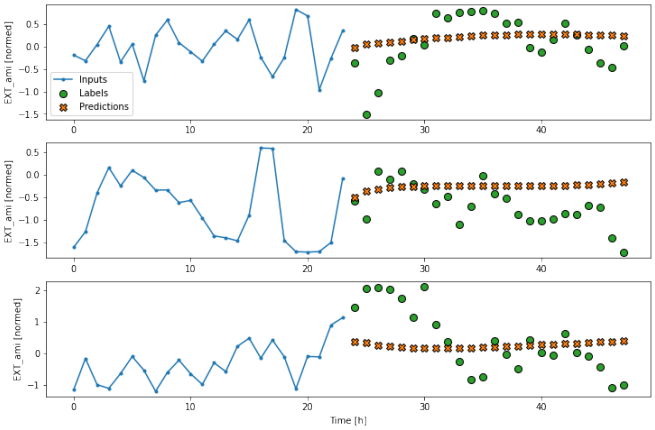

그림 47: LSMT 다중 예측 결과

제 5 장

결론 및 고찰

5.1 결론 및 요약

데이터 기반의 LDPE MI 예측을 위해 다양한 기계학습 모델을 적용해 본 결과 분류 방식 모델인 XGboost에서 기존 운전원의 정성적 판단을 대신해 공정 물성관리에 활용 가능한 수준의 성능을 확인했다.

모델 구현은 우선 데이터 전처리 과정부터 시작되며 전처리로 이상반응 및 설비 고장에 의한 이상치를 제거했다. 이 과정은 공정 전문가의 개입 및 판단이 필요하며 고성능 모델 구현에 가장 큰 영향을 끼친다.

데이터 전처리 실시 후 다중 회귀 모델에 사용될 독립변수를 선택하기 위해 실제 공정 운전에 사용되는 독립변수 17개를 선별하여 MI의 상관관계를 분석했다. 분석 결과 AMI값을 제외한 공정 데이터는 강한 상관관계를 보이지 않았으나 변수 유의성 판단을 위한 회귀 분석 결과 17개 독립변수 중 12개의 독립변수가 유의한 것을 확인하여 최종 17개, 12개로 구분하여 모델에 반영하였다.

최초 수립한 다중 회귀 모델의 경우 R^2_{adj}값이 30.6%로 설명력이 매우 낮았다. 그러나 공정 데이터를 MI 측정 간격과 동일한 측정간격으로 구분한 이후 구간 평균을 이용해 수립한 모델은 설명력이 각각 45.5%로 1.5배 가까이 상승함을 알 수 있었다. 전자의 경우 샘플링 시점의 공정 데이터를 독립변수로 이용했으나 MI 측정값은 샘플링 시점의 값이 아닌 연속 생산 공정에서 샘플링 전후 구간의 대표값을 의미하므로 구간 평균을 이용하는

것이 바람직함을 알 수 있다. 구현된 모델은 AMI로 MI를 예측하는 단순 회귀 모델의 설명력 34.1% 대비 45.5%로 34% 개선된 성능을 보였으나 절대적 설명력이 여전히 낮아 현업에 사용하기는 성능이 부족하다.

그러나 정확한 MI값 예측이 아닌 정품 여부 분류만으로도 운전자의 물성 조절 판단에 도움을 줄 수 있고 이러한 분류 모델의 효용성을 고려해 다양한 분류 모델 성능을 AMI의 MI 예측 수준을 기준으로 비교했다. MI 데이터는 불량품 데이터의 비중이 1.4%로 극단적으로 양품이 많다. 한쪽으로 치우친 이진(binary) 데이터 분류 모델 성능을 정확히 비교하기 위해서 정확도와 함께 재현율을 이용했다.

현재 사용하고 있는 AMI의 MI 예측 성능은 정확도 0.953, 재현율 0.270이며 한쪽으로 치우친 데이터로 높은 정확도 대비 재현율이 낮음을 알 수 있다. 분류 모델의 성능 개선을 위해 하이퍼 파라미터 조정을 거쳐 최종 모델 학습을 완료하였고 여러 분류 모델 중 나이브 베이즈, 배깅, XGboost 모델을 제외하면 재현율 기준 reference 대비 유사하거나 낮은 성능을 보였다. 이는 데이터 불균형이 영향을 주기 때문이다.

분류 모델 중 XGBoost 모델은 모델 학습에 사용되는 다양한 하이퍼 파라미터를 임의로 조정할 수 있으며 R에서 제공하는 베이지안 최적화로 간단하게 최적화할 수 있다. 개선된 XGBoost 모델의 성능은 정확도 0.969, 재현율 0.556으로 reference 대비 재현율 기준 206% 향상됐다. 기존 AMI 값을 대체하여 공정 물성 관리에 충분히 사용할 수 있는 수준이다.

마지막으로 MI 데이터가 가진 시계열 특징에 착안하여 시계열 분석을 실시하였다. 시계열 분석은 여러 종류의 공정 데이터 없이 MI값 만으로 분석이 가능한 장점이 있다. 우선 ARIMA 분석으로 결정된 ARIMA(0,1,1) 모델로 값을 예측하였으나 예측값 신뢰구간(-0.20 -+0.18)이 MI 실제 스펙 구간과 유사하여 현업에는 사용 불가능한 수준이었다. LSTM 모델은 학

습이 잘 되지 않아 성능이 MAE 기준 0.160으로 reference인 AMI의 50% 수준이었다. 하지만 MI 데이터 수가 부족하여 학습이 원활치 않아 발생한 현상이었으며 17만개의 AMI 데이터를 이용한 시계열 예측에서는 MAE 0.017으로 우수한 성능을 확인할 수 있었다. 추후 충분한 데이터가 확보된다면 분류 모델인 XGBoost와 함께 현업 활용이 가능한 성능이다.

5.2 한계 및 후속연구

자율주행 분야에서 괄목할 만한 인공지능 실적을 보여주고 있는 자동차 산업과[19] 달리 국내 석유화학산업의 성과는 여전히 미비하다. 그럼에도 본 연구는 대표 석유화학 제품인 LDPE 용융지수를 기계 학습으로 예측해 기존 사용중인 정성적 방법을 뛰어넘는 예측 성능을 확인했다.

하지만 이번 연구의 한계점으로 실제 현장에서 사용되는 독립변수를 사용해 예측과 분류 모델을 구현하였음에도 예측 모델 설명력이 50% 미만이었으며, XGBoost를 제외한 일부 분류 모델에서도 기대 이하의 성능을 보였다. 설명변수의 질적, 양적 결함이 원인이며 무엇보다 모델에 사용된 공정 데이터와 종속변수의 낮은 상관관계 개선이 필요하다. 이를 위해 첫번째, 기존 설치 센서들의 수집 간격을 일치시켜 독립변수의 데이터 전처리 과정을 간소화하고 센서 데이터의 검교정을 강화해 데이터 품질을 향상시켜야 한다. 두번째, 연구 대상을 전 공정으로 확대해 모델 예측 성능을 향상 시킬 가능성 있는 독립변수를 추가 발굴해야 한다. 연구 대상 공정을 압축 공정을 비롯한 전 공정으로 확대해 압축기 토출 압력, 온도 반응기 입구 조건 같은 독립변수를 모델 학습에 추가로 사용하고 종속변수의 수집 기간을 확대해 전반적인 데이터의 양(volume)을 확대하여 성능이 부족했던 모델을 재검증해야 한다.

그러나 이런 한계에도 대표적 석유화학 공정에서 예측, 분류, 시계열의 다양한 방법론 유용성을 검증 했으며, LDPE 용융지수를 기존 AMI 대비 재현율 기준 최대 206% 향상된 성능으로 예측했다. 본 연구 모델은 실시간 센서 데이터와 연동시켜 실제 산업현장의 용융지수 예측에 활용할 계획이다.

참고 문헌

[1] 김주영 and 조진환, "제조업 인력 고령화와 정년연장," *Issue Paper 2012-291*, 2012.

[2] I. Kadel, T. Herrmann, and M. Busch, "Modeling free radical ethylene polymerization with multifunctional comonomers in tubular reactors: influence on microstructural ldpe properties," in *Macromolecular Symposia*, vol. 324, pp. 67–77, Wiley Online Library, 2013.

[3] K. Yadav and R. Thareja, "Comparing the performance of naive bayes and decision tree classification using r," *International Journal of Intelligent Systems and Applications*, vol. 11, no. 12, p. 11, 2019.

[4] P. A. Jaskowiak, R. Campello, *et al.*, "Comparing correlation coefficients as dissimilarity measures for cancer classification in gene expression data," in *Proceedings of the Brazilian symposium on bioinformatics*, pp. 1–8, Brasília Brazil, 2011.

[5] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 18, no. 6, pp. 275–285, 2004.

[6] Y. Freund, R. E. Schapire, *et al.*, "Experiments with a new boosting algorithm," in *icml*, vol. 96, pp. 148–156, Citeseer, 1996.

[7] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[9] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[10] W.site: https://otexts.com/fpp3/arima.html, ”Forecasting:Principles and Practice”. Chapter9 ARIMA models, ”Accessed 24September2022”.

[11] W.site: https://colah.github.io/posts/2015-08-Understanding-LSTMs/. ”Understanding LSTM Networks”, ”Accessed 24September2022”.

[12] D. M. Powers, “Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation,” *arXiv preprint arXiv:2010.16061*, 2020.

[13] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[14] W.site: https://www.r-bloggers.com/2022/01/using-bayesian-optimisation-to-tune-a-xgboost-model-in-r/. ”Using bayesian optimisation to tune a XGBOOST model in R”,”Accessed 24September2022”.

[15] W.site: https://gist.github.com/JunmoNam/a87f107d064178956e614f9940c55b92. ”JunmoNam/xgboost.r”, ”Accessed 24September2022”.

[16] P. C. Phillips and P. Perron, “Testing for a unit root in time series regression,” *Biometrika*, vol. 75, no. 2, pp. 335–346, 1988.

[17] S. Hochreiter and J. Schmidhuber, *Long short-term memory*. Neural Computation, 1997.

[18] W.site: https://www.tensorflow.org/tutorials/structured_data/time_series?hl=en. ”Time series forecasting”, ”Accessed 24September2022”.

[19] B.-Y. Lee, “국내외 자율주행자동차 기술개발 동향과 전망,” *Information and Communications Magazine*, vol. 33, no. 4, pp. 10–16, 2016.

Abstract

LDPE melt flow index prediction study based on machine learning model

Hangrae Kim
Graduate School of Practical Engineering
Seoul National University

Low density polyethylene melt index control is executed by the operator determining the management direction based on the process sample analysis result and reflecting it to the process operating conditions. However, in the time period when there is no sample analysis, the operator predicts and adjusts the melt index by referring to the inter-process melt index measurement value after analyzing the process situation, and as a result, the distribution of physical properties of the entire product is determined accordingly. In this study, a melt index prediction study based on a machine learning model using process data unrelated to proficiency was conducted to improve product quality deviations caused by operator proficiency.

The temperature and pressure data collected in the process were used as independent variables, and the melt index analyzed and recorded in the laboratory was preprocessed in two forms, continuous and discrete, respec-

tively, to be used in prediction and classification models. The predictive performance of multiple regression analysis using continuous data is based on R^2, RMSE, and MAE as evaluation indicators, and classification models such as XGBoost using discrete data use accuracy, precision, and recall as evaluation indicators. The performance improvement rate compared to the simple melt index analyzer was compared. In addition, the possibility of using deep learning technology was verified by predicting the melt index with time series analysis tools of ARIMA and LSTM.

This study has verified the usefulness of prediction, classification, and time-series methodologies in representative petrochemical processes, and is meaningful in presenting machine learning methodologies for various petrochemical processes that consider melt index as a major property.

Keywords : LDPE, machine learning, melt index, prediction of properties, quality management

Student Number : 2021-21941