

산업용 보일러 효율에 영향을 미치는 주요 인자 탐색

최문석¹ 윤새봄¹ 김기범¹ 한승훈¹

¹경북대학교

Exploring the Key Factors Affecting Industrial Boiler Efficiency

Choi Mun Seok¹ Yoon Sae Bom¹ Kim Ki Bum¹ Han Seong Hoon¹

¹Kyung Pook National University

요약

보일러는 산업 현장의 핵심 설비로 본 연구에서는 IoT 시스템을 통해 수집한 보일러 가동 데이터를 통해 보일러의 효율에 영향을 미치는 주요 인자를 파악하려고 한다. 주요 인자를 파악하는 것은 효율 개선을 통해 저탄소 배출과 에너지 절약효과를 기대할 수 있다. PCA를 통해서 보일러효율에 영향을 미치는 주성분을 분석한다. 그 후 보일러 작동 상태 변화와 관련된 첫 번째 그리고 두 번째 주성분을 제거하여 보일러 작동상태에서의 효율 영향인자를 파악할 수 있게 한다. XGBoost Regressor 모델을 사용하여 학습을 진행하고 SHAP(Shapley Additive exPlanations)를 활용해 주요 인자를 파악한다.

1. 서론

보일러는 산업 현장의 핵심 설비로 다양한 영향이 효율에 영향을 미치고 있다. 산업용 보일러의 효율은 KS B 6205라는 엄격한 KS 규정에 따라 측정된다. 본 연구에서는 국내 소재 산업용 보일러 회사가 본 규정에 따라 개발한 효율 계산 시스템과 IoT 시스템을 통해 수집한 보일러 가동 데이터를 통해 가동상태에서 보일러의 효율에 영향을 미치는 효율 인자를 파악하려고 한다. 산업용 보일러 효율에 미치는 인자들을 바탕으로 효율 개선 방향을 마련하여 효율향상을 기대할 수 있다. 이에 따라 저탄소 배출과 에너지 절약 효과를 누릴 수 있다. 더불어 이를 바탕으로 원칙대로 작동되는 보일러를 기대할 수 있을 것으로 기대된다. Wenbia Wang, Lan Chen, Zhanyuan Ge and Si yuan Wang (2014)는 보일러의 운영 효율을 향상하기 위해 주성분 분석을 활용한 관리 모델을 제안했다. 보일러의 운영 효율이 감소하면, 영향을 미치는 매개변수를 기여도 그래프를 통해 파악할 수 있다. 본 논문에서 사용하는 데이터의 보일러 효율 값인 ‘열 손실법 효율’ 계산식은 LNG 성분, 공급 가스 관련 값과 운전 데이터값으로 복잡하게 계산된다. 따라서 본 논문에서는 기계학습 모델인 XGBRegressor를 사용하여 변수 간 비선형적 관계를 파악한다. 성능이 좋게 학습된 모델은 데이터의 특징이나 패턴, 관계 등을 잘 파악하여 예측에 활용하고 있다고 할 수 있다. 그리고 보일러 효율은 여러 입력 변수 간의 복잡한 상호 작용에 의해 결정된다. 따라서 비선형 관계를 모델링 하는 것은 이런 데이터의 특성을 잘 파악해 주요 인자를 더 효율적으로 식별할 수 있다.

2. 이론적 배경

2.1 주성분 분석(PCA; Principal Component Analysis)

주성분 분석은 고차원 데이터의 주요 특성을 파악하고 단순화하기 위한 방법이다. 이 방법은 원본 데이터에서 가장 큰 변동성을 갖는 축(주성분)을 찾아 데이터를 새로운 좌표계로 변환한다. PCA는 데이터의 차원을 줄여 계산의 복잡성을 감소시키고, 중요한 정보를 유지하면서 노이즈를 제거하는 데 유용하다.

2.2 XGBoost Regressor

XGBoost Regressor는 회귀 분석에 주로 사용되는 부스팅 앙상블 기법의 최신 발전 형태로, Gradient Boosting의 진보된 버전으로 볼 수 있다. 이는 약한 학습자인 결정 트리를 순차적으로 학습하여 강력한 예측 모델을 구축한다. Regularization 기법을 도입하여 과적합을 효과적으로 제어하며, 학습률, 트리의 깊이, 정규화 등의 하이퍼파라미터를 조절하여 모델의 일반화 성능을 높일 수 있다.

2.3 베이지안 최적화

베이지안 최적화(Bayesian Optimization)는 함수의 최적값을 찾기 위해 확률적 모델을 사용하여 검색 공간을 탐색하는 최적화 방법이다. 주로 머신러닝 모델의 하이퍼파라미터 튜닝 등에 활용되며, 목적 함수의 평가가 비용이 많이 드는 경우에 특히 효과적이다. 베이지안 최적화의 기본 단계는 다음과 같다. 1) 사전 확률 분포 설정: 목적 함수에 대한 초기 불확실성을 나타내는 사전 확률 분포를 설정한다. 가우시안 프로세스(Gaussian Process, GP)가 사용되는데, 가우시안 프로세스는 다음과 같이 정의된다.

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

2) 가우시안 프로세스 활용: GP는 목적 함수의 사전 분포로 사용되며, 관측된 데이터를 기반으로 함수에 대한 확률적 추정을 제공한다. 3) 취득 함수(Acquisition Function) 선택: 취득 함수는 다음 평가 지점을 선택하는 데 사용된다. 이 함수는 불확실성이 크고 예상되는 목적 함수값이 좋은 지점을 선호한다. 취득 함수 중 하나인 ‘Expected Improvement’는 다음과 같이 정의된다.

$$EI(x) = E[\max(f(x) - f(x^+), 0)]$$

4) 최적화 및 업데이트: 취득 함수를 최적화하여 다음 평가 지점을 선택하고, 새로운 관측값으로 가우시안 프로세스를 업데이트한다. 5) 반복: 원하는 정도의 최적화가 될 때까지 3번과 4번 단계를 반복한다.

2.3 SHAP

SHAP(Shapley Additive exPlanations)은 기계학습 모델의 예측을 이해하고 각 특성이 예측에 미치는 기여를 설명하는 기술로, Shapley 값의 원리를 적용한다. Shapley 값은 게임 이론에서 플레이어 간의 기여를 공정하게 나누는 개념이며, SHAP은 이를 모델 해석에 적용하여 각 특성의 중요도를 정량화한다. 모델 예측에 대한 SHAP 값은 해당 특성이 예측을 얼마나 변화시키는지를 나타내며, 이를 통해 개별 예측의 해석이 가능하게 한다.

3. 분석결과

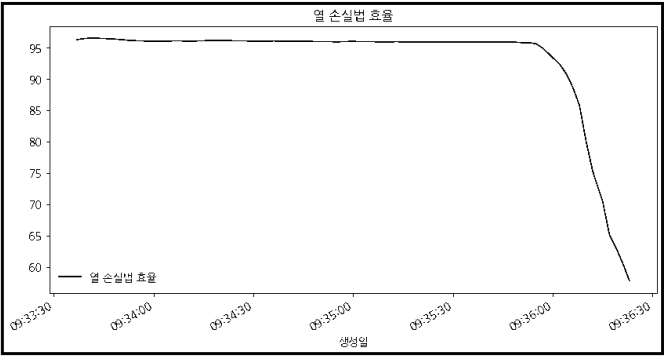
3.1 데이터 전처리

본 연구에서는 A 공장에서 가동중인 국내 소재 보일러 회사의 노동연관식 산업용 보일러에서 측정된 데이터를 사용한다. 09월 01일부터 10월 31까지의 데이터를 사용하였다. 산업용 보일러는 항상 가동되지 않고 가동과 중지를 반복한다. 따라서 보일러가 작동되지 않는 상태에서는 효율 값이 0이다. 따라서 본 연구에는 보일러가 작동되는 상태의 데이터만 뽑아 수직 합병하여 사용한다. 모든 값이 일정하거나 모든 값이 0인 독립변수 그리고 종속변수와 관련 없는 독립변수는 제거하여 사용하였다. 총 34개의 독립변수와 종속변수인 ‘열 손실법 효율’로 분석을 진행하였다.

번호	변수명	번호	변수명
1	부하율	18	배기가스 온도3
2	보일러 압력	19	배기 재 순환 온도
3	송풍기 인버터 출력	20	에코 온도1
4	송풍기 입력	21	에코 온도2
5	급수 펌프	22	배기가스 NOx
6	급수 펌프 입력	23	배기가스 O2
7	가스 댐퍼	24	재순환 O2
8	가스 댐퍼 입력	25	급수량(적산유량)
9	Air 댐퍼	26	급수량(순간유량)
10	Air 댐퍼 입력	27	연료량(적산유량)
11	재순환 외기 댐퍼	28	연료량(순간유량)
12	재순환 댐퍼 입력	29	소비전류
13	재순환 외기 댐퍼 입력	30	진동센서1
14	급수 수위	31	진동센서2
15	보일러 온도	32	송풍기 고장확률
16	배기가스 온도1	33	AIR 댐퍼 고장확률
17	배기가스 온도2	34	GAS 댐퍼 고장확률

<표1. 독립변수 종류>

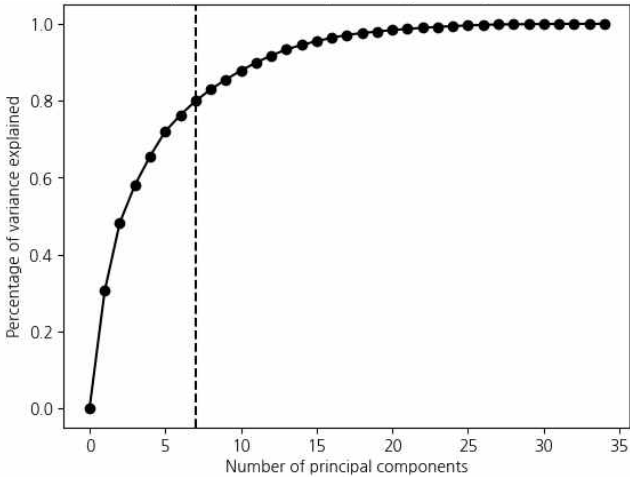
그림1과 같이 보일러의 열 손실법 효율은 보일러 중단 시점이 다가올 때 급격하게 감소하는 모습을 볼 수 있다.



<그림1. 작동되는 상태의 열 손실법 효율>

3.2 주성분 분석(PCA; Principal Component Analysis)

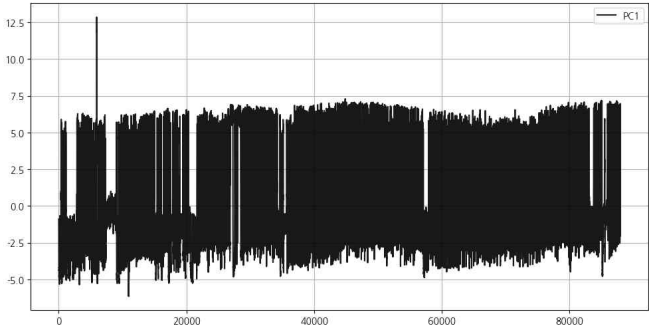
데이터의 주요한 특징이나 패턴을 추출하기 위해서 주성분 분석을 진행한다. 우선 보일러 전체 데이터에 대해 Z-score 정규화를 진행한다. 주성분 개수는 누적 분산 설명률의 80%를 설명하는 부분에서 결정한다. 따라서 그림1과 같이 전체 분산의 79.9%를 설명하는 7개의 주성분을 선정한다.



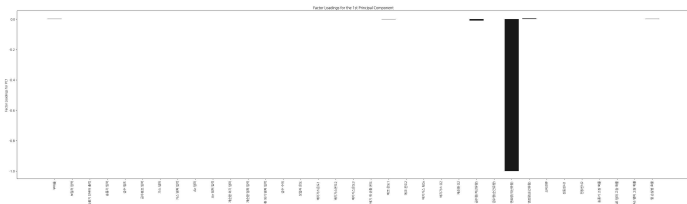
<그림2. 주성분 개수에 따른 분산 설명력>

3.2.1 첫 번째 주성분

그림3과 같이 보일러 작동상태에서 시간에 따른 첫 번째 주성분의 경과를 파악할 수 있다. 첫 번째 주성분의 베타 로딩을 그림4에서 파악할 수 있는데, 해당 주성분은 대부분 ‘연료량’ 변수로 이루어져 있다는 것을 알 수 있다.



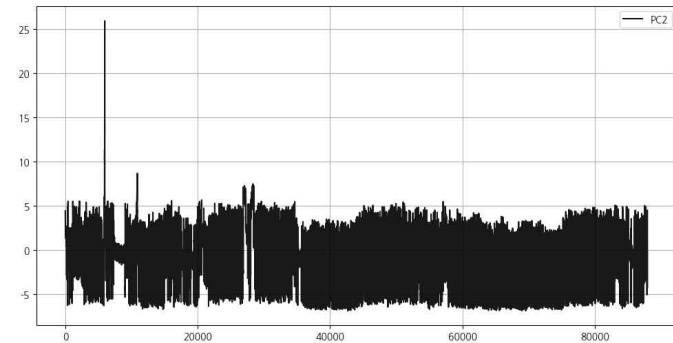
<그림3. 첫 번째 주성분>



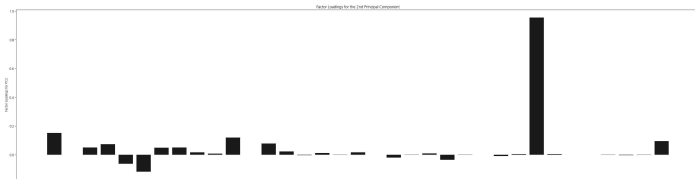
<그림4. 첫 번째 주성분의 베타 계수>

3.2.2 두 번째 주성분

그림5와 같이 보일러 작동상태에서 시간에 따른 두 번째 주성분의 경과를 파악할 수 있다. 두 번째 주성분의 베타 로딩을 그림6에서 파악할 수 있는데, 해당 주성분은 첫 번째와 마찬가지로 대부분 ‘연료량’ 변수로 이루어져 있다는 것을 알 수 있다.



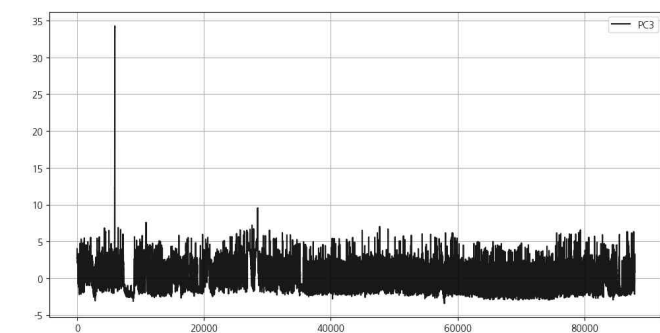
<그림5. 두 번째 주성분>



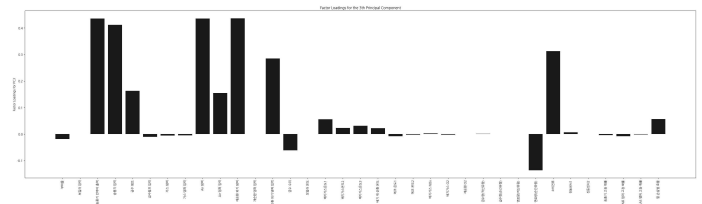
<그림6. 두 번째 주성분의 베타 계수>

3.2.3 세 번째 주성분

그림7과 같이 보일러 작동상태에서 시간에 따른 세 번째 주성분의 경과를 파악할 수 있다. 세 번째 주성분의 베타 로딩을 그림8에서 파악할 수 있는데, 해당 주성분을 구성하는 주요 변수로는 차례로 ‘송풍기 인버터 출력’, ‘재순환 외기 댐퍼’, ‘송풍기 입력’, ‘소비전류’ ‘재순환 외기 댐퍼 입력’ 이다. 따라서 세 번째 주성분은 보일러 시스템의 공기량 관련 요인이라고 볼 수 있다.



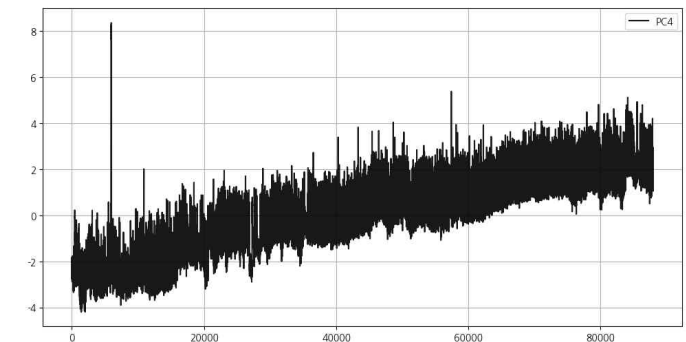
<그림7. 세 번째 주성분>



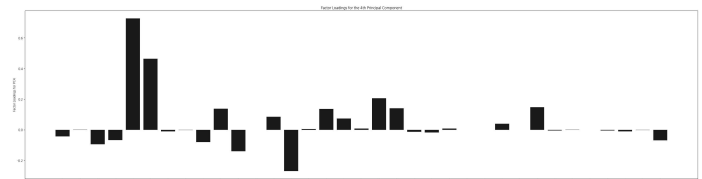
<그림7. 세 번째 주성분의 베타 계수>

3.2.4 네 번째 주성분

그림8과 같이 보일러 작동상태에서 시간에 따른 네 번째 주성분의 경과를 파악할 수 있다. 네 번째 주성분의 베타 로딩을 그림9에서 파악할 수 있는데, 해당 주성분을 구성하는 주요 변수로는 차례로 ‘급수펌프’, ‘급수펌프 압력’ 이다. 따라서 네 번째 주성분은 보일러 시스템의 물의 양 관련 영향이라고 볼 수 있다. 다만, 이전의 주성분과는 다르게 장기적인 추세를 가지는데, 이는 보일러 시스템이 가동되면서 장기적으로 누적되는 효과가 있거나 외생변수에 보일러 시스템이 영향을 받고 있음을 의미한다.



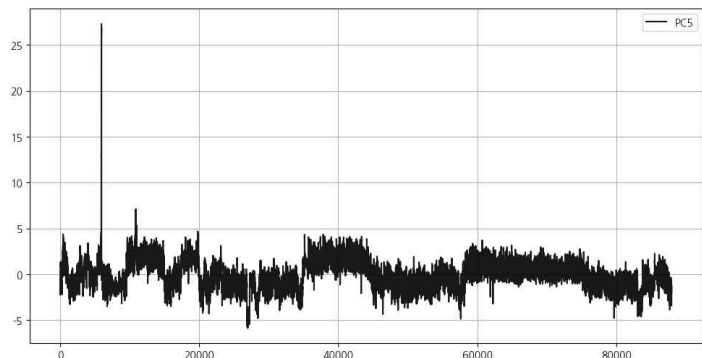
<그림8. 네 번째 주성분>



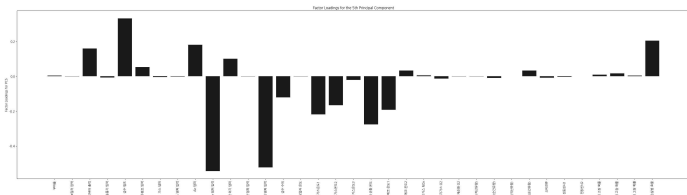
<그림9. 네 번째 주성분의 베타 계수>

3.2.5 다섯 번째 주성분

그림10과 같이 보일러 작동상태에서 시간에 따른 다섯 번째 주성분의 경과를 파악할 수 있다. 다섯 번째 주성분의 베타 로딩을 그림11에서 파악할 수 있는데, 해당 주성분을 구성하는 주요 변수로는 차례로 ‘Air 댐퍼 입력’, ‘재순환 외기 댐퍼 입력’ 이다. 따라서 다섯 번째 주성분은 보일러 시스템의 공기의 양 관련 영향이라고 볼 수 있다. 다만, 네 번째 주성분과 마찬가지로 장기적으로 특정 요인에 반응하는 모습을 보인다.



<그림10. 다섯 번째 주성분>



<그림11. 다섯 번째 주성분의 베타 계수>

여섯 번째와 일곱 번째 주성분은 큰 특징이 없어 따로 언급하지 않겠다.

3.3 주요 인자 탐색

3.3.1 주성분 제거

앞선 주성분 분석 결과, 첫 번째 그리고 두 번째 주성분은 거의 ‘연료량’ 변수로 이루어져 있음을 알 수 있었다. 보일러가 가동대기 상태로 들어갈 때, 즉 연료의 공급이 중지될 때 그림1과 같이 열효율의 변화가 압도적으로 크다. 따라서 첫 번째 그리고 두 번째 주성분을 보일러 가동 상태 변화라고 볼 수 있다. 본 연구의 목적은 보일러 작동 상태에서의 효율 영향인자를 찾는 것으로, 첫 번째 그리고 두 번째 주성분을 제거한다. 데이터를 주성분 공간으로 변환하고, 첫 번째 그리고 두 번째 주성분을 제외하여 데이터의 원래 형태로 복원시켜 해당 과정을 수행한다.

3.3.2 모델 훈련 및 테스트

정규화 및 주성분 제거된 데이터를 8:1:1 비율로 각각 훈련, 검증, 평가 데이터로 분할한다.

가. XGBRegression 모델의 최적 하이퍼 파라미터 선정
본 연구에서는 XGBoost 모델의 하이퍼 파라미터 튜닝을 위해서 베이지안 최적화 기법을 사용¹⁾한다. 하이퍼 파라미터 탐색공간은 표2와 같다. 최적화에 훈련과 검증 데이터를 사용하였으며 손실함수는 RMSE를 사용하였다.

하이퍼 파라미터 변수명	탐색공간
max_depth	트리의 최대 깊이를 1에서 5까지 1씩 증가시키며 탐색한다.
gamma	0에서 1사이의 값을 균등하게 탐색한다.
reg_alpha	0에서 50사이의 값을 균등하게 탐색한다.
reg_lambda	10에서 100사이의 값을 균등하게 탐색한다.
colsample_bytree	0에서 1사이의 값을 균등하게 탐색한다.
min_child_weight	0에서 5사이의 값을 균등하게 탐색한다.
n_estimators	10000
learning_rate	0에서 0.15사이의 값을 균등하게 탐색한다.
max_bin	200에서 550까지 1씩 증가시키며 탐색한다.
random_state	5

<표2. 베이지안 최적화 탐색공간>

베이지안 최적화 결과 ‘max_depth’는 5, ‘gamma’는 약 0.00723, ‘reg_alpha’는 약 35.6084, ‘reg_lambda’는 약 16.9919, ‘colsample_bytree’는 약 0.7352, ‘min_child_weight’는 약 4.9031, ‘learning_rate’는 약 0.1373 그리고 max_bin은 539로 최적화되었다.

나. XGBRegression 모델 학습 및 테스트

앞서 찾아낸 최적의 하이퍼 파라미터 값을 바탕으로 XGBRegression모델을 학습²⁾을 진행하였다. 훈련 및 평가 데이터에서 진행한 모델의 성능은 표3과 같다.

데이터	성능지표	값
훈련 데이터	RMSE	1.5405
	MAPE	1.3796
평가 데이터	RMSE	1.6596
	MAPE	1.4473

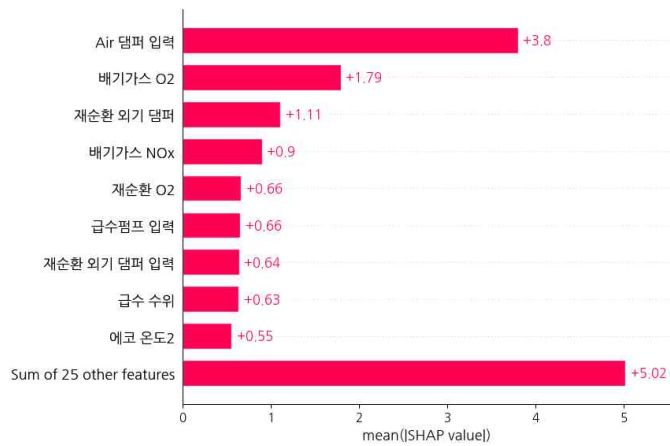
<표3. 모델의 성능>

3.4 SHAP(Shapley Additive exPlanations)

샤플리 값을 분석한 글로벌 중요성은 그림12와 같다. Air 덤퍼 입력, 배기가스 O₂, 재순환 외기 덤퍼, 배기가스 NO_x, 재순환 O₂, 급수펌프 입력, 재순환 외기 덤퍼 입력, 급수 수위, 예코 온도2 순으로 중요도가 높은 것을 확인할 수 있다.

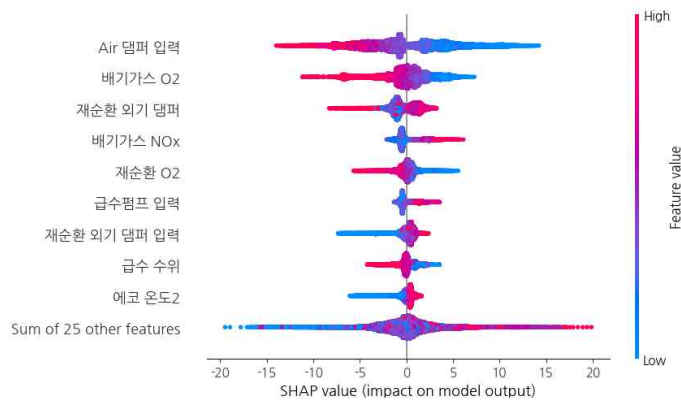
1) Google Colab이 제공하는 Tesla T4 GPU를 사용하여 수행되었습니다.

2) 로컬 환경에서 Intel I5 CPU를 사용하여 수행되었습니다.



<그림12. 특성 중요도>

형상 중요도와 형상 효과를 결합해 그림13에 SHAP Summary Plot을 나타내었다. SHAP 피쳐 의존도와 상호작용은 부록에 첨부하였다.



<그림13. SHAP Summnay Plot>

4. 결 론

본 연구에서는 주성분 분석을 통해 보일러효율에 영향을 미치는 주성분을 분석하였다. PCA 결과 전체 분산의 약 30.5%를 설명하는 첫 번째 주 성분과 전체분산의 약 17.8%를 설명하는 두 번째 주성분을 보일러 상태변화와 관련된 요인으로 파악했다. 본 연구는 보일러 작동 상태에서의 효율 영향인자를 찾는 것으로 두 주성분을 제거해 사용한다. 베이지안 최적화 과정을 통해 최적화된 XGBoost Regressor 모델을 사용하여 비선형 관계를 모델링하였고 SHAP(Shapley Additive exPlanations)를 활용해 주요 인자를 파악하였다.

‘Air’ 댐퍼 입력, ‘배기가스O2’와 같은 주요 인자들의 영향력과 순위가 그림 12에 나타난다. 해당 결과는 열공학적 보일러 효율 영향인자와 일치하는 결과로, 해당 분석이 실제 보일러 가동 상태를 잘 반영하였다.

5. 참고문헌

[1] Wenbiao Wang, Lan Chen, Xinjie Han, Zhanyuan Ge, and Siyuan Wang. (2014)Modeling and Application of Principal Component Analysis in Industrial Boiler. ISNN, Advances in Neural Networks, ISNN 2014, 395-404

6. 부록

6.1 shap 값 산점도

