RESEARCH ARTICLE

건물 에너지 데이터 분석에서 결측치 처리방식에 따른 차원 축소 및 모델 예측 성능 비교

이경재[1] · 임현우[2†]

[1]건국대학교 일반대학원 건축학과 건축공학전공, 석사과정
[2]건국대학교 건축대학 건축학부, 조교수

Comparing the Performance of Dimensional Reduction and Model Prediction Performance Due to Missing Data Handling for Building Energy Data Analysis

Lee Kyungjae[1] · Lim Hyunwoo[2†]

[1]Master Course in Architectural Engineering, Department of Architecture, Graduate School of Konkuk University
[2]Assistant Professor, Department of Architecture, College of Architecture, Konkuk University

[†]Corresponding author: hyunwoolim@konkuk.ac.kr

Abstract

Handling missing values during data analysis is an important issue that directly affects the prediction performance of models and research results. However, research on the differences between the dimensionality reduction rate and model prediction performance is still lacking for building energy-related data. Therefore, this study compared the dimensionality reduction rate and model prediction performance by handling missing values in weather information datasets, which is related to building energy. The missing value-handling methods were divided into removal, k-nearest neighbors (KNN) imputation, and no handling. Dimensionality reduction methods were classified based on principal component analysis and feature selection using the model. Further, the eXtreme Gradient Boosting (XGBoost) algorithm, a gradient boosting method with its own missing data handling capabilities, was used. Consequently, few principal components were required to explain 95% of the variance in the raw data when the missing values were removed than when they were replaced with KNN. Moreover, the dimensionality reduction methods of model building and feature selection outperformed principal component analysis in terms of dimensionality reduction rate and model predictive accuracy. Particularly, the XGBoost model without missing values had the highest accuracy, suggesting that the missing-value handling method of XGBoost may be superior to conventional missing-value handling methods. These results may have important implications for selecting imputation methods in building energy data analysis, considering the effort and cost of missing value handling, and can significantly reduce the cost and effort of data preprocessing.

Keywords: XGBoost 모델(XGBoost), 기계학습(Machine learning), 주성분분석(Principal component analysis), 기상정보(Weather data), 결측치 처리방법(Missing values handling methods)

Journal of the Korean Solar Energy Society
Vol.44, No.1, pp.59-75, February 2024
https://doi.org/10.7836/kses.2024.44.1.059

pISSN : 1598-6411

eISSN : 2508-3562

Received: 22 October 2023

Revised: 23 November 2023

Accepted: 26 December 2023

기호 및 약어 설명

XGBoost : 익스트림 그레디언트 부스팅 알고리즘(eXtreme Gradient Boosting)
KNN : K-최근접 이웃 알고리즘(K-Nearest Neighbor)
RMSE : 평균제곱근 오차(Root Mean Square Error)
CVRMSE : 평균제곱근 오차의 변동계수(CV, Coefficient of Variation)
R^2 : 결정계수(R-squared)
PCA : 주성분 분석(Principal Component Analysis)
EDA : 탐색적 데이터 분석(Exploratory Data Analysis)
SHAP : 섀플리 분석(SHapley Addictive exPlanations)

그리스 기호 설명

r : 피어슨 상관계수(Pearson Correlation Coefficient)

1. 서 론

건물 에너지 데이터의 분석과 활용은 지속가능한 사회와 건물 에너지 효율성 향상에 필수적인 요소이다. 특히 건물 에너지 사용패턴에 대한 이해와 예측은 국가적 측면에서의 에너지 소비 최적화, 절약 전략을 개발하는데 중추적인 역할을 한다. 그러나 이러한 데이터 분석 작업에는 언제나 '결측치 처리(Missing values handling)' 문제가 동반된다. 데이터 분석에서 결측치 처리는 모델의 예측 성능과 연구 결과에 직접적인 영향을 끼치는 중요한 문제이며, 건물과 관련된 다양한 종류의 데이터 또한 결측치 처리 문제에서 자유로울 수 없다. 대표적으로 건물 에너지 소비와 직접적인 관련이 있는 기상정보의 경우 결측치 비율이 상당하며, 이를 어떻게 처리하고 분석할지에 대한 결정은 건물 에너지 평가 모델의 예측 성능과 데이터 해석력에 중대한 영향을 미친다.

가장 많이 사용되는 결측치 방식은 대표적으로 결측치 제거(Removal), 단일 대체(Single imputation), 다중 대체(Multiple imputation) 방법인 모델 기반 대체 세 가지가 있다[1]. 결측치 제거방식의 경우 가장 단순하지만 분포의 왜곡과 정보의 과도한 제거가 문제가 된다. 단일 대체의 경우 대표적으로 평균대체와 중앙값 대체가 있으며, 앞선 방식과 마찬가지로 변수의 분포를 심하게 왜곡할 위험이 있다. 모델 기반 대체방식은 관찰된 데이터를 바탕으로 추정된 여러 개의 사후 분포에서 도출된 값으로 결측값을 대체한 뒤, 완성된 여러 데이터 집합을 생성하고 각 데이터 집합을 활용하여 Rubin[2]이 개발한 방법을 통해 추정값을 결합하는 방법이다. 이러한 모델 기반 대체는 앞선 두 방식보다 결측치를 더 잘 대체하긴 하지만 계산시간이 많이 걸린다는 단점이 있다. 이렇듯 결측치 대체 방식에 대한 선택은 결측치의 종류, 데이터의 특성, 분석 목적에 따라 달라지며, 더 나은 결측치 대체 방식에 대해서는 지속적으로 연구되고 있다.

최근 기계학습 관점에서 결측치 처리에 대한 새로운 대안으로 XGBoost를 활용한 결측치 처리 방식이 주목받고 있다. XGBoost는 워싱턴 대학교의 Chen and Guestrin[3]이 개발한 익스트림 그레디언트의 줄임말로, 모델

구축전 결측치를 대체하거나 따로 처리하지 않아도 자체적으로 결측치를 처리하여 분석할 수 있는 능력을 갖추고 있어, 전통적인 결측치 처리 방식과 비교되는 새로운 가능성을 제공한다. 이에 관한 연구로는 Mustika et al.[4]의 생명보험 가입자의 위험수준을 측정하는 문제에서 Bayesian Ridge, Random forest, decision tree 같은 결측치 처리방식 보다 XGBoost의 결측치 처리 방식이 더 좋은 성능을 보였으며, Rusdah and Murfi[5]의 연구에서는 결측치 제거(Removal), k-nearest neighbors (KNN) 대체보다 XGBoost의 결측치 처리 방식이 평균적으로 더 높은 정확도를 보였다.

하지만 이러한 기존 연구는 분류문제에만 한정되어 있으며, 주로 life insurance 데이터에서 XGBoost의 결측치 처리 성능에 대해서만 연구했다. 따라서 아직까지 건물 에너지 분야에서 예측문제와 관련한 XGBoost의 결측치 처리방식의 효용성에 대해서는 연구가 부족한 상황이다. 이에 본 연구는 결측치를 제거하거나 다른 방법으로 대체하는 대신, XGBoost를 이용하여 결측치를 처리하지 않은 채로 예측 모델을 구축하는 방법의 유용성을 검토하고자 한다. 데이터는 기상정보와 건물 에너지 사용량 데이터를 활용하였으며, 건물에너지 사용량 예측에 있어 결측치 처리방식과 차원 축소 전략이 모델 성능에 어떠한 영향을 미치는지를 중점적으로 파악하였다. 결측치 처리방식은 결측치 제거, KNN대체, 결측치 무처리 방식을 사용하였으며 차원 축소 전략은 주성분분석(Principal component analysis, 이하 PCA)과 모델 구축을 통한 특징 선택으로 나누어 비교하였다. 이를 통해 기존의 결측치 처리 방식과 XGBoost를 통한 접근법이 건물 에너지 예측 모델의 성능에 어떠한 차이를 만들어 내는지 규명함으로써 건물 에너지 데이터 분석에 새로운 방법론을 제시하고자 한다.

2. 방법론

Fig. 1은 본 연구의 포괄적인 흐름과 사용된 방법들을 나타낸다. 먼저 전국의 기상 정보를 수집하여 데이터셋을 생성하고, 이 데이터셋의 결측치에 무처리, 제거(Removal) 그리고 KNN 결측치 대체 방법을 적용하였다. 다음 단

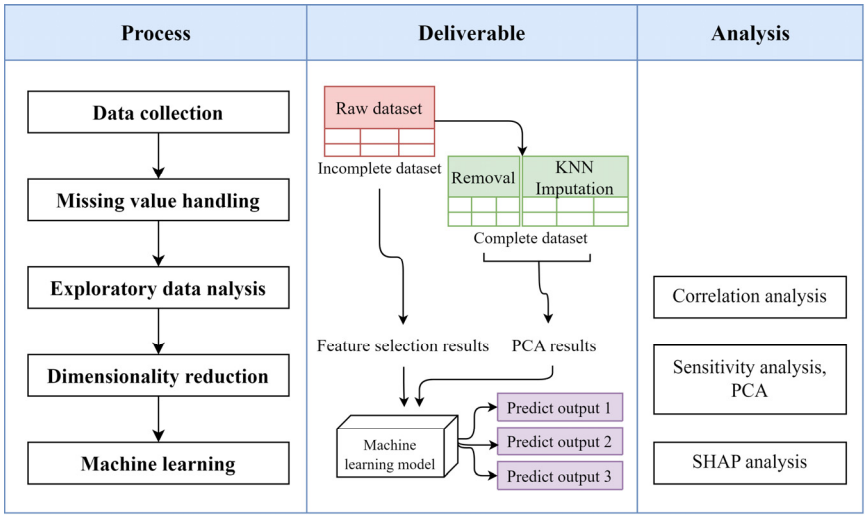


Fig. 1 Procedure of research

계에서는 각각의 결측치 처리 방법이 데이터에 어떠한 영향을 미치는지 파악하기 위해 탐색적 데이터 분석(Exploratory data analysis, 이하 EDA)을 실시하였다. 그 후 데이터의 차원 축소를 모델 구축을 통한 특징 선택(Feature selection) 방식과 PCA방식으로 나누어 진행한 뒤, 각 차원 축소 방식에 따른 모델의 정확도를 비교하였다.

2.1 데이터 수집 및 정제

기상정보 데이터는 기상청의 관측 원점 월 기상정보를 활용하였으며, 해당 정보는 기상청 기상자료개방포털[6]에서 다운로드 할 수 있다. 본 연구에서는 2018년 1월부터 2022년 12월까지의 월별 기상정보를 사용하였으며, 이 데이터는 총 1,244,920의 행과 19개의 변수로 구성되어 있다. 이 중 10개 변수는 기상과 관련된 정보를 담고 있으며 나머지 9개 변수는 관측 지점의 위치정보가 기재되어 있다. 본 연구에서는 기상정보 내 기상과 관련된 10개의 변수에 대해서만 데이터셋을 제작해 분석을 진행했으며 해당 변수는 Table 1에 기재되어 있다.

평균해면기압(Average sea level pressure)은 관측소에서 관측한 기압을 관측소의 해발고도에 맞게 보정한 기압이다. 합계일조시간(Total sunshine hours)은 한 달 동안 측정된 일조시간의 총량을 나타내며, 합계일사량(Total solar radiation)은 한 달 동안 태양에서 오는 태양복사에너지가 지표에 닿는 양을 나타낸다. 소형총증발량(Small total evaporation)과 대형총증발량(Large total evaporation)은 각각 소형/대형 증발접시에 담긴 물의 전날 관측값에 대한 차이값의 한 달 총량을 나타낸다. Table 1에서 확인할 수 있듯이 소형/대형총증발량의 결측치가 46.97%로 가장 높으며, 합계일사량 또한 46.77%로 결측치가 높다. 이러한 결측치가 발생한 원인은 특정 기상지점에서 소형총증발량과 대형총증발량을 측정하지 않는 경우가 있으며, 주로 증발량을 측정하지 않은 관측지점에서 일사량도 측정 하지 않는 경우가 많기 때문이다. 나머지 변수들은 결측치가 거의 없으며, 이러한 결측치 비율은 기상정보개방포털에서 제공하는 원데이터의 결측치를 그대로 반영한 것이다. 또한 모든 변수의 데이터 타입은 숫자형이며 연속적이다.

Table 1 Test bed information

Parameter	Example	Unit	Missing value rate (%)
Average temperature	-2	°C	0.01
Average sea level pressure	1025.5	hPa	0.03
Monthly precipitation	20	mm	0.12
Average relative humidity	61	percent (%)	0.05
Total sunshine hours	165.9	hour (hr)	0.00
Sunlight rate	55.02	percent (%)	0.00
Total solar radiation	229.95	MJ/m^2	46.77
Average wind speed	1	m/s	0.15
Small total evaporation	38.6	mm	46.97
Large total evaporation	27.3	mm	46.97

건물 에너지 사용량은 서울시 강남구, 마포구, 영등포구의 22143개 모든 건물의 전기와 가스에너지를 합산한 서울시 3개구 월별 전체 에너지 데이터를 활용했다. 이상치 제거 방식은 사분위수(Interquartile range, 이하 IQR)범위에서 크게 벗어난 median ±1.5 × IQR 값을 제외하는 3IQR방식을 활용했다.

2.2 결측치 (Missing value)

결측치는 데이터셋 내에서 어떤 이유로든 값이 없는 경우를 말한다. 결측치는 데이터 분석과 모델링에서 중요한 문제로 작용하는데, 결측치를 전부 제거하면 막대한 정보의 손실을 일으킬 수 있고, 결측치를 잘못 대체하면 데이터의 편향이 생길 수도 있기 때문이다. 따라서 데이터 분석에서 결측치 처리는 분석가의 노하우와 견해가 가장 많이 반영되는 동시에 분석 결과에 결정적인 영향을 미치는 중요한 과정이다.

결측치는 종류에 따라 크게 세 가지로 구분된다[2]. 첫째, 완전 무작위 결측(Missing completely at random, MCAR)은 결측치간에 아무런 상관성이 없는 것으로, 데이터를 입력한 이가 실수했거나 전산상의 오류가 난 경우가 있다. 둘째, 무작위 결측(Missing at random, MAR)은 결측치가 다른 변수와 상관성이 있는 경우이다. 예를 들어 특정 집단(노인)이 다른 집단(청년)보다 어떤 질문에 답하지 않을 확률이 높은 경우 MAR로 분류될 수 있다. 마지막으로 앞선 두 경우에 속하지 않는 비무작위 결측(Missing Not at Random, MNAR)이 있다. MNAR은 결측치가 발생한 이유가 관측된 것뿐만이 아닌 관측하지 못한 어떤 변수와 관련이 있으며 사실상 결측치 처리가 가장 어려운 경우이다. 예를 들어 소득수준을 파악하고자 하는 연구에서 소득이 매우 높은 사람들은 소득 정보를 기입하지 않는 경향이 있어, 결측치가 발생하는 원인이 소득수준과 직접적으로 관련이 있는 상황이라면 이는 MNAR에 해당한다.

결측치 처리에 관해서는 다양한 연구가 존재하며 데이터의 종류에 따라서도 결측치 처리방식이 달라진다. Akande et al.[7]의 연구에서는 범주형 데이터에 대한 다중 대체 방식에 관해 연구했으며, Zhang et al.[8]과 Ma and Chen[9]의 연구에서는 베이지안 접근 방식을 활용한 Markov Chain Monte Carlo (MCMC) 샘플링을 통해 결측치를 다루었다. Dewi et al.[10]의 연구에서는 결측치를 0 (zero), 평균, 중앙값, 같은 열의 데이터의 최빈값 등으로 대체하여 처리하는 방법을 연구했다. 건축 공학 부문에서는 Kim et al.[11]의 태양광 발전에 대한 예측에 관한 연구에서 KNN대체를 활용하여 결측치를 다룬 연구가 있다.

(1) KNN 대체(KNN Imputation)

KNN 대체는 데이터셋 내 관측치 간의 유사성을 활용하여 결측치를 추정하고 대체하는 방법이다[12]. 결측치를 가진 데이터 포인트를 예측하려는 대상으로 보고, 해당 데이터 포인트와 가장 가까운 유클리드 거리(식(1))를 가지는 k개의 이웃 데이터를 찾아 이 이웃들의 평균(연속형 변수일 경우) 또는 최빈값(범주형 변수일 경우)을 사용하여 결측치를 대체하는 알고리즘이다.

$$distance(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + \ldots + (z_2 - z_1)^2} \quad (1)$$

2.3 상관관계 분석

본 연구에서는 입력변수 간의 상관성을 피어슨 상관계수를 통해 파악했다[13]. 피어슨 상관계수는 칼 피어슨이 개발한 두 변수 간의 선형 연관성을 측정하는 상관계수로 상관관계 분석시 스피어만(spearman)분석과 함께 많이 쓰이는 방법이다. 피어슨 상관계수는 −1에서 +1사이의 값을 가지며 +1은 완벽한 양의 상관관계, 0은 상관관계 없음, −1은 완벽한 음의 상관관계를 의미한다. 피어슨 상관 계수는 두 변수간의 공분산을 표준 편차의 곱으로 나누어 구하며 공식은 식(2)과 같다.

$$r = \frac{cov(X, Y)}{s(X) \times s(Y)} \quad (2)$$

2.4 피쳐 엔지니어링

피쳐 엔지니어링은 머신 러닝의 성능을 향상하기 위해 데이터를 변환, 선택, 개선하는 모든 프로세스를 의미한다. 그중 차원 축소는 데이터의 정보를 최대한 유지하면서 변수를 줄이는 기법이다. 데이터의 차원이 커질수록 계산 비용이 증가하고, 모델이 복잡해져 과적합의 위험이 커진다. 이러한 차원의 저주를 해결하기 위해 차원 축소를 진행한다.

차원 축소의 방법들 가운데 대표적으로 피쳐 추출과 피쳐 선택이 있다. 피쳐 추출은 기존 피쳐를 변형하거나 조합하여 새로운 피쳐를 생성하는 방법으로 주로 주성분 분석을 사용한다. 반면, 피쳐 선택 방식은 데이터셋의 피쳐 중에서 분석 목적에 중요한 피쳐만을 선택하는 방법으로 민감도 분석, 상관분석 등과 같은 방식을 활용하여 변수를 선택한다.

(1) 주성분 분석(Principal component analysis, PCA)

주성분 분석은 데이터셋의 원본 특성을 유지한 채 소수의 주성분으로 데이터셋을 압축시키는 방법이다[14]. 이 방법은 주로 데이터 내 변수간의 다중공선성을 해결하고, 데이터셋을 간결하게 구성하여 차원의 저주 문제를 해결하기 위해 사용된다. 대표적인 방법으로는 PCA (Principal component analysis), KPCA (Kernal principal component analysis), LDA (Linear discriminant analysis)등이 있으며 본 연구에서는 비지도 선형 변환 기법인 PCA 방법을 사용하였다. 주성분 개수의 기준은 연구에서 통상적으로 사용하는 원데이터 분산의 95%를 설명하는데 필요한 주성분의 수를 기준으로 하였다[14].

(2) 민감도 분석

민감도 분석은 모델의 출력변수에 대한 입력변수의 영향도를 정량화하는 방법으로, 국부 민감도 분석과 전역 민감도 분석(Global sensitivity analysis)으로 나눌 수 있다[15]. 국부 민감도 분석은 입력변수의 한 값을 변경하였을 때 모델의 출력변수에 미치는 영향도를 분석하는 방법이고, 전역 민감도 분석은 입력변수의 모든 값을 변화시키면서 모델의 출력변수에 미치는 영향도를 분석하는 방법이다. 변수의 개수가 많거나 모델이 복잡해질수록, 국부 민감도 분석만으로는 입력변수의 민감도를 정확히 평가하기 어렵기 때문에 전역 민감도 분석 방법이 권장된다. 대표적인 전역 민감도 분석 방식으로는 분산기반 방식의 Sobol 방법과, Morris 방법이 있다[16].

본 연구에서는 최근 주목받고 있는 설명가능한 AI 기반의 전역 민감도 분석방법인 SHAP를 활용하였다[16,17]. SHAP는 새플리 값(Shapley value)을 활용하여 변수의 영향력을 평가하는 방법이다. 기존 민감도 분석 방법은 변수의 중요도를 정량화 하는데 그치는 반면 SHAP 민감도 분석은 입력변수가 출력변수에 미치는 영향력의 크기와 방향성을 보다 직관적으로 파악할 수 있다.

2.5 머신러닝

본 연구에서 활용된 XGBoost 알고리즘은 트리 기반의 앙상블(Ensemble) 모델로 기존의 그래디언트 트리 부스팅 모델을 개선한 버전이다. 2016년 Chen et al.[3]이 개발한 이 모델은 다른 기계학습 알고리즘과 다르게 결측치를 처리하는 자체적인 알고리즘을 가지고 있으며, 이 기능은 결측치가 많은 데이터셋에서 유용하게 활용할 수 있다. XGBoost의 결측치 매커니즘은 다음과 같다.

결측치가 있을 때 XGBoost는 각 결측치를 트리의 어느 방향(오른쪽 노드, 왼쪽 노드)으로 보낼지 학습하는 과정을 진행한다. 트리의 각 노드에서 이러한 과정을 진행하면서 어느 방향으로 가는 것이 목표 변수를 더 잘 예측하는지 판단하는데 이러한 과정에서 손실 함수를 계산하여 손실이 최소화되는 방향으로 결측치를 보낸다. 모델이 학습된 후, 새로운 데이터에 적용할 때 만약 변수의 값이 결측치라면, 학습 과정에서 학습된 방향으로 결측치를 보내 예측을 진행한다. 이러한 방식은 XGBoost가 결측치를 처리할 필요 없이 원데이터의 결측치를 자체적으로 처리할 수 있게 해주며, 이는 데이터 전처리 과정에서의 결측치 처리 비용을 생각하면 매우 효율적이다.

(1) 모델 학습 및 평가

기상정보는 다양한 기상 변수들로 구성되어 있으며, 각 변수의 스케일과 단위가 다르다. 이러한 상대적 크기 차이로 인하여 모델 구축 시 큰 스케일을 가진 변수들이 모델에서 더 큰 영향력을 갖게 되고, 작은 스케일의 변수들은 상대적으로 무시될 수 있는 위험성이 있다. 따라서 모델 학습 전 변수들을 동일한 스케일과 단위로 맞춰주는 작업인 스케일링 작업이 필요하다. 본 연구에서는 최대-최소 정규화를 활용하여 모든 변수 값을 [0,1]로 동일한 범위로 만들었다(식(3)). 여기서 X는 원래 값이며, $Xmin$, $Xmax$는 각각 데이터의 최소/최대값, $Xnorm$은 정규화된 값이다. 이를 통해 모든 변수를 동일한 범위로 만들어 모델 구축 시 변수의 상대적 크기 차이에 따

른 영향을 최소화하였다.

$$Xnorm = \frac{X - Xmin}{Xmax - Xmin} \tag{3}$$

전체 데이터에서 모델에 학습시킬 학습 데이터와 모델의 성능을 테스트할 테스트 데이터는 7:3으로 나누었다. 모델의 성능을 최적화하기 위해서는 하이퍼파라미터 튜닝 과정이 필수적이다. 이를 위해 학습데이터 또한 4:1의 비율로 나누어 하이퍼파라미터 튜닝 과정을 진행하였다. 하이퍼파라미터 튜닝 방법들에는 대표적으로 Grid Search, Random Search, Bayesian Optimization 방식이 있다[18]. 본 연구에서는 미리 정의된 범위 내의 하이퍼파라미터의 모든 가능한 조합을 시도하여 최고의 성능을 보이는 하이퍼파라미터의 조합을 선택하는 Grid Search 방식을 활용하였다. 하이퍼파라미터 튜닝 범위는 Table 2와 같다.

모델 성능 평가 지표는 결정계수(R-squared)와 Coefficient of variation of the root mean squared error (CVRMSE) 활용하였다. R^2은 회귀분석에서 예측의 적합도를 측정하는 통계치로 모델이 얼마나 실제 데이터를 잘 설명하는지에 대한 정보를 제공해준다. R^2은 0에서 1사이 값으로 나타나며 종종 백분율로 표현되기도 한다. R^2은 식(4)를 통해서 구해진다.

$$R^2 = 1 - \frac{SSres}{SStot} \tag{4}$$

여기서 $SSres$는 잔차 제곱합, 즉 예측된 값과 실제값 차이의 제곱 합을 나타내고, $SStot$는 전체 데이터의 평균과 실제값 차이의 제곱합을 나타낸다. R^2이 1에 가까울수록 모델이 데이터를 완벽하게 예측하는 것이며 R^2이 0에 가까울수록 데이터를 예측하지 못한다.

CVRMSE 또한 모델의 예측 성능을 평가하는 데 사용되는 통계지표 중 하나로 Root mean squared error (RMSE)를 평균값으로 나누어 정규화한 값이다. CVRMSE는 예측값과 실제값 간의 오차를 백분율로 표현하며 식(5)를 통해서 계산된다.

$$CVRMSE = \left(\frac{RMSE}{\overline{Y}}\right) \times 100\% \tag{5}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2} \tag{6}$$

여기서 $\overline{Y}$는 실제 값 Y의 평균이며, $\hat{Y}_i$는 i번째 예측값, Y_i는 i번째 예측값이다. 식(6)에서 RMSE는 모델 예측 값과 실제 값의 차이 제곱합의 평균으로 계산되며, CVRMSE는 이 값을 평균으로 나누어 상대적인 오차를 얻는다. 그 후, 이 값을 백분율로 표현하여, 모델의 오차가 평균에 대해 얼마나 큰지를 비율로 표현한다. 즉 CVRMSE가 낮을수록 모델의 예측 성능이 좋다고 판단할 수 있다.

Table 2 XGBoost hyperparameter tunning range

Hyperparameter	Range
Max depth	1-8
Estimators	10-200
Min child weight	0.5-1
Subsample	0.5-1
Colsample by tree	0.7-1
Colsample by node	0.6-1
Colsample by level	0.6-1

3. 연구 결과

3.1 결측치 처리방식에 따른 상관관계 분석 결과

기존 데이터셋은 1,244,920개의 행으로 구성되었으나, 결측치를 제거하는 과정에서 658,997개의 행으로 축소되었다. 이는 결측치를 제거하는 과정에서 데이터의 약 절반이 손실되었음을 의미한다. 따라서 이로 인한 정보의 손실과 분석결과의 신뢰성 하락에 주의해야 한다. 각 결측치 처리 방법에 따른 변수 간의 상관관계 분석 결과는 Fig. 2, Fig. 3을 통해 확인할 수 있다.

상관관계 분석 결과, 결측치 처리 방법에 상관없이 변수간의 상관관계 분석 결과는 일관된 패턴을 보였다. 특히 평균온도와 평균해면기압 간의 상관관계는 결측치 처리 방식과 관계없이 뚜렷한 음의 상관관계(r = −0.92)를 보였으며, 이는 온도와 기압 간의 기존 지식과 일치하는 결과이다. 소형총증발량과 대형총증발량 간에는 상관관계는 완벽한 양의 상관관계(r = +1)를 보였으며, 이는 모델 구축 시 다중공선성 문제를 야기할 수 있으므로 해당 변수의 제거 또는 결합을 고려해야 한다. 한편 주목할 점은 결측치를 제거한 데이터셋에서 합계일사량과 소형/대형총일사량간의 매우 높은 상관관계가(r = +0.94) 나타났다는 점이다. 이는 결측치를 대체한 데이터셋에서 관측된 상관관계(r = +0.77)보다 상당히 높은 결과이며, 이는 결측치 처리방식이 단순히 데이터의 양에만 영향을 미치는 것이 아니라, 변수간의 관계에도 상당한 영향을 미칠 수 있음을 시사한다.

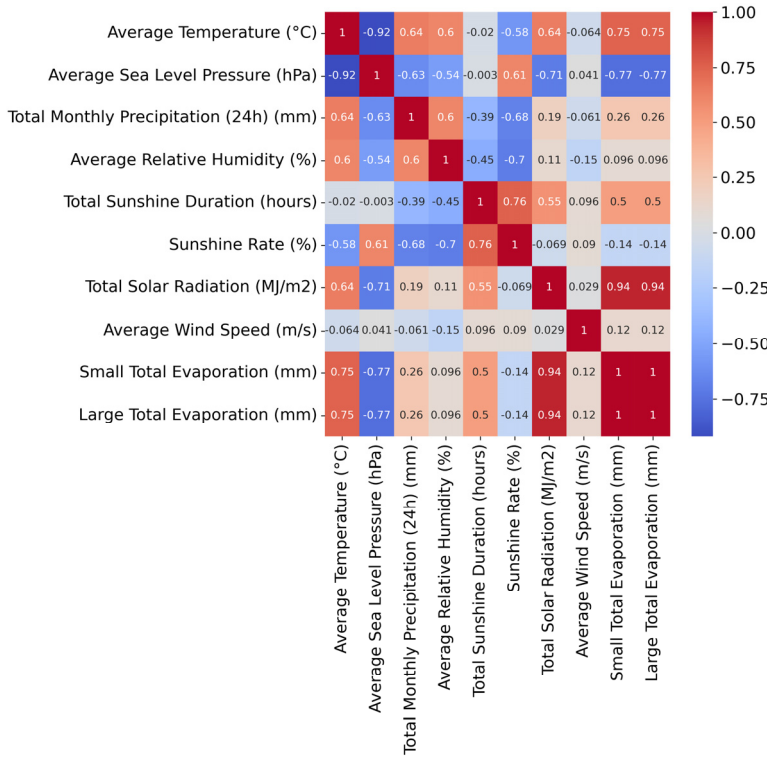


Fig. 2 After missing value removal correlation results

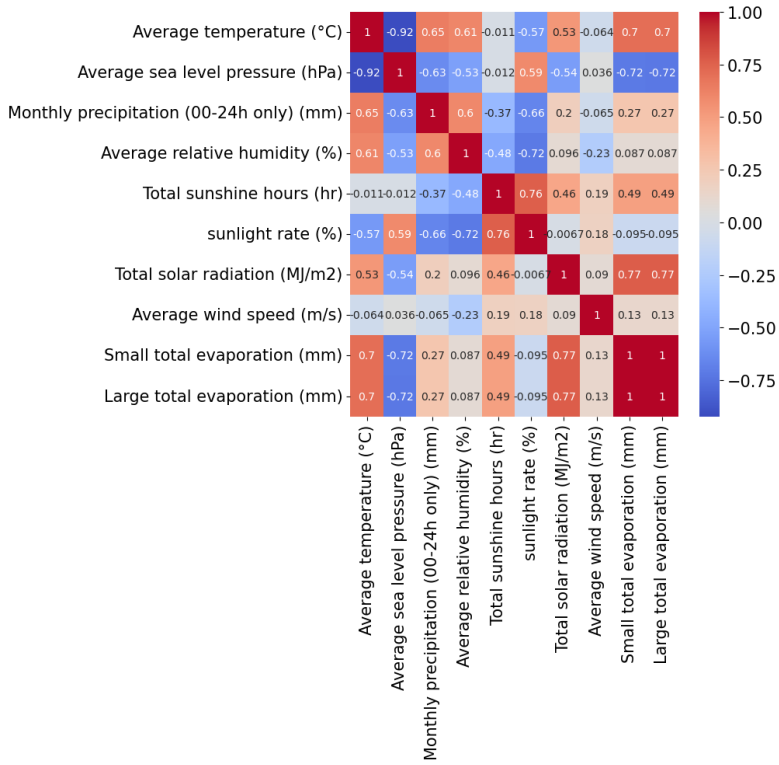
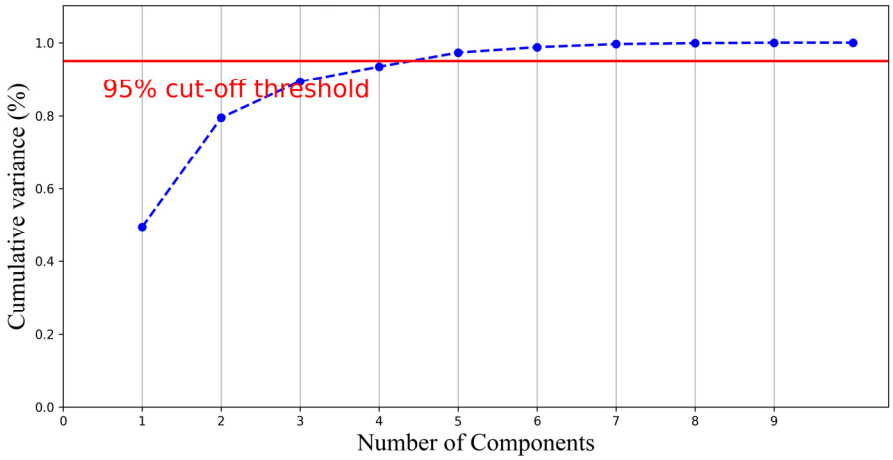


Fig. 3 After missing value imputation correlation results

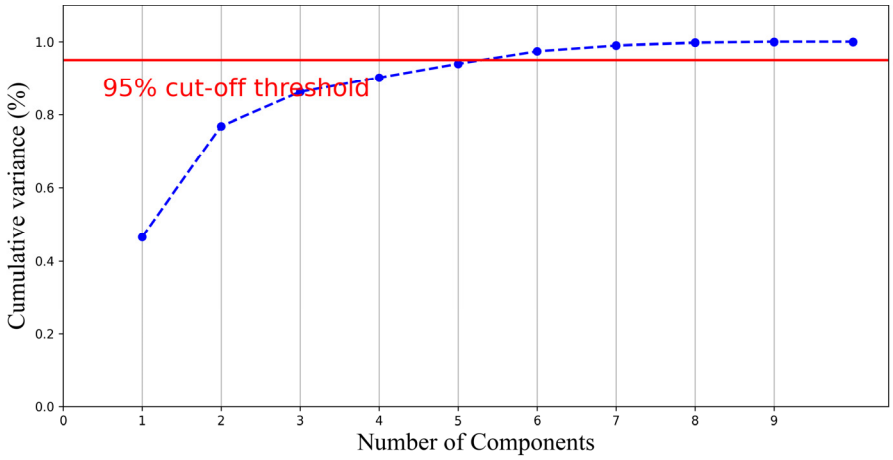
3.2 결측치 처리방식에 따른 PCA결과 비교

Fig. 4은 결측치 처리방식에 따른 PCA 결과를 나타낸 그래프이다. Fig. 4를 통해 결측치 처리 방식에 따라 원데이터 분산의 95%를 설명하기 위해 필요한 주성분의 개수가 차이가 있음을 확인할 수 있다. 결측치를 제거한 데이터셋은 5개의 주성분으로 원데이터 분산의 대부분을 설명할 수 있지만, 결측치를 대체한 데이터셋은 6개의 주성분이 필요하였다.

이는 결측치 제거 과정에서 일부 정보가 손실되어 데이터의 분산구조가 단순화되었을 가능성을 나타낸다. 또한 3.1절에서의 상관관계 분석 결과에서 결측치를 제거한 기상정보 데이터셋의 합계일사량과 소형/대형총증발량간의 상관관계(r= +0.94)가 결측치를 대체한 경우의 상관관계(r= +0.77)보다 높게 나타났는데, 이는 때문에 결측치를 제거한 경우에서 분산의 구조가 단순해지고 따라서 필요한 주성분의 개수가 줄어들었다고 해석할 수 있다. 이러한 결과는 앞선 상관관계 분석 결과와 마찬가지로 결측치 처리 방식의 선택에 따라 데이터의 구조와 분석 결과가 달라질 수 있음을 시사한다.



(a) PCA results of missing value removal



(b) PCA results of missing value imputation

Fig. 4 PCA results due to missing values handling method

3.3 피쳐 선택(Feature selection) 결과

기상정보의 피쳐 선택을 위해 XGBoost를 활용하여 회귀 모델을 제작했다. 모델의 입력값은 서울 지역의 결측치가 있는 기상정보로 각각의 변수와 결측치 비율은 Table 1과 같다. 타겟값은 강남구, 마포구, 영등포구의 에너지 사용량을 합산한 2018-2022년 월별 전체 건물에너지 사용량이다. 모델 하이퍼파라미터 튜닝 결과는 Table 3과 같다. 학습된 모델의 예측 성능은 테스트 데이터인 18개월의 에너지 사용량을 기준으로 R^2 = 0.87의 성능을 보였다(Fig. 5). 이는 ASHRAE Guidline 14[19]에서 제시한 기준보다 우수한 성능으로 피쳐 선택을 위한 모델로 활용하기 적합한 모델 성능이다.

Table 3 Best XGBoost hyperparameter for no missing value handling

Hyperparameter	Value
Max depth	1
Estimators	20
Min child weight	2
Subsample	0.6
Colsample by tree	1
Colsample by node	0.8
Colsample by level	0.8

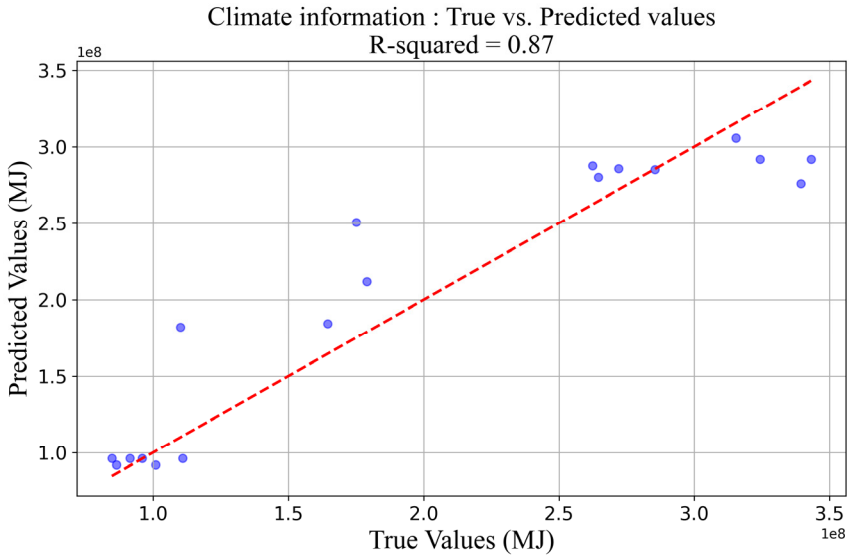


Fig. 5 Comparison between model predict value and true value

제작한 모델을 활용하여 SHAP 민감도 분석을 진행하였으며, Fig. 6은 입력데이터에 대한 SHAP 값의 절대값을 평균 내어 나타낸 막대그래프이다. 분석 결과 전체 에너지 사용량에 미치는 변수의 중요도는 평균기온, 소형

총증발량, 평균상대습도, 평균해면기압 순으로 중요했다. 이 결과를 바탕으로, 중요도가 낮은 변수를 하나씩 제거하는 방식으로 피쳐 선택을 수행해, 최적의 변수 조합을 탐색했다.

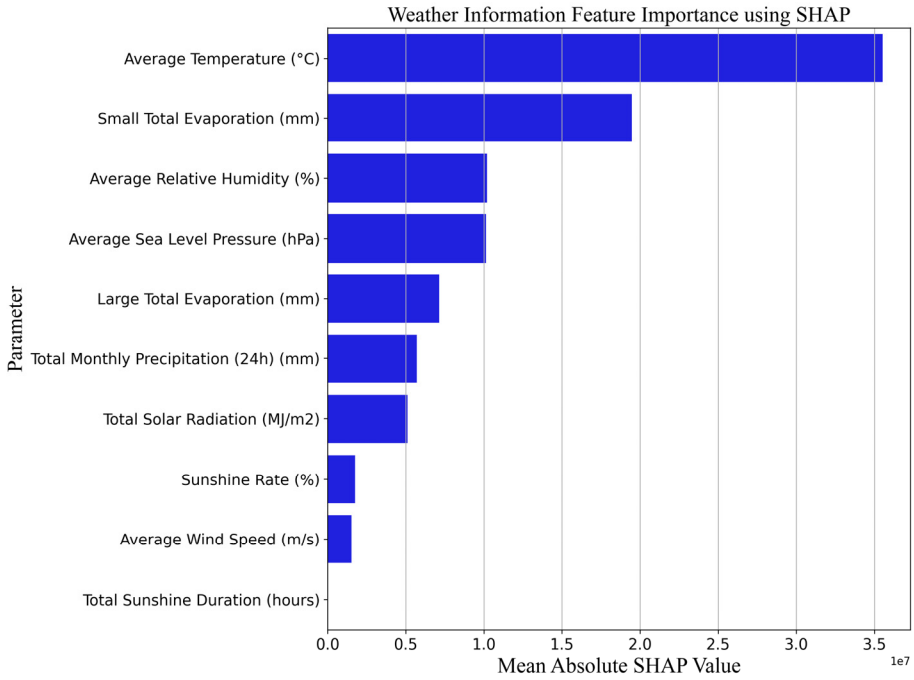


Fig. 6 SHAP analysis results

Fig. 7은 변수 개수에 따른 모델의 성능을 CVRMSE로 측정한 그래프이다. 처음에는 변수의 개수가 늘어날수록 모델의 성능이 좋아졌으나, 변수 4개 이상부터는 변수의 증가가 모델의 예측 성능을 더 이상 향상시키지 못하였다. 특히, 변수의 개수가 6개 이상인 지점부터 모델의 정확도는 CVRMSE 약 14% 수준으로 수렴하는 경향을 보였다. 이는 변수의 추가가 일정 수준 이상에서는 모델 성능 측면에서 추가적인 이득을 제공하지 않는다는 점을 의미한다. 최적의 변수 조합은 평균기온, 소형총증발량, 평균상대습도의 조합으로 나타났으며 이 3개의 변수를 사용하여 모델을 구성할시 CVRMSE 9.57%의 가장 좋은 성능을 보였다. 따라서 피쳐 선택의 결과로 이 최적의 변수 3개를 선정하였으며, 이는 기존 10개의 변수 중 3개의 주요 변수만 선택하여 차원을 70% 감축한 결과로 모델의 계산 효율성과 필요한 데이터의 감축 측면에서 긍정적인 결과이다.

또한, 이러한 결과는 이전의 PCA를 사용하여 데이터셋의 차원을 줄인 결과와 비교하여 더 많은 차원을 줄일 수 있음을 보여주는 결과이다. 결측치를 제거한 데이터셋에서는 PCA를 사용하여 50%의 차원 감소율을 보였고, 결측치를 대체한 데이터셋에서는 40%의 차원감소율을 보인 반면, 피쳐 선택은 70%의 차원을 축소하였기 때문이다.

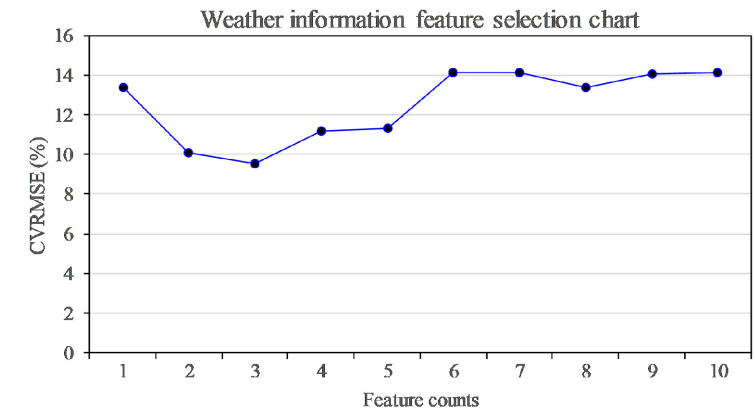


Fig. 7 Model performance due to the number of variable

4. 차원축소 방식에 따른 모델 정확도 비교

차원축소 방식에 따른 모델의 정확도 차이를 파악하기 위해 추가로 2개의 모델을 구성하여 총 3개의 차원축소 방식에 대한 모델 정확도를 비교 분석하였다. 비교 대상이 되는 차원축소방식은 다음과 같다. 첫째, 결측치를 제거한 데이터셋의 PCA방식, 둘째 결측치를 KNN 대체한 PCA방식, 마지막으로 모델 기반의 피쳐 선택을 통한 차원축소 방식이다.

모델 구축시 입력값으로 PCA 방식에서는 전체 분산의 95%를 설명하는 데 필요한 주성분의 최소조합을 입력값으로 사용하였고, 모델 기반의 피쳐 선택에서는 이전에 선택된 세 변수(평균기온, 소형총증발량, 평균상대습도)를 활용하였다. 모델 학습 시 진행된 데이터 분할, 하이퍼파라미터 튜닝, 목표 출력값은 모두 이전의 피쳐 선택 시 활용한 모델과 동일하게 적용하였다. 상세한 하이퍼파라미터 튜닝 결과는 Tables 4, 5에 기재되어 있다.

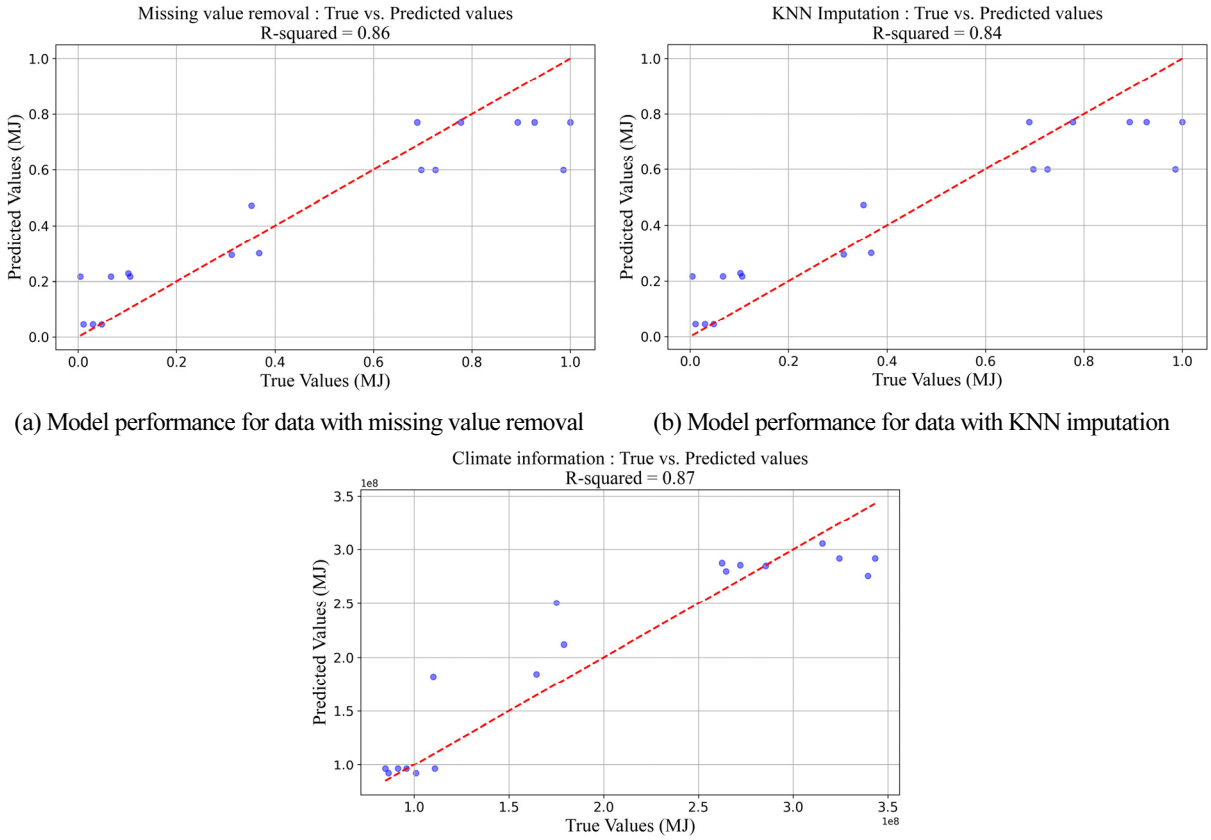
Fig. 8은 테스트데이터를 기준으로 모델 정확도를 비교한 Scatter plot이다. 모델 정확도 비교 결과 결측치를 처리하지 않은 XGBoost 회귀 모델이 R^2 = 0.87로 가장 높은 성능을 나타냈다. 이에 비해, 결측치를 제거한 경우와 KNN 대체한 경우는 각각 R^2 = 0.86, R^2 = 0.84의 성능을 보였다. 모델간의 정확도 차이가 크지 않지만 이러한 결과는 XGBoost를 활용할 시 결측치를 처리하지 않고도 우수한 성능의 모델을 제작할 수 있다는 것을 의미한다. 이는 특히, 기존 모델구축 시 결측치 처리에 드는 비용을 생각하면 매우 흥미로운 결과이다.

Table 4 Best XGBoost hyperparameter for data with missing value removal

Hyperparameter	Value
Max depth	1
Estimators	10
Min child weight	1
Subsample	1
Colsample by tree	1
Colsample by node	1
Colsample by level	1

Table 5 Best XGBoost hyperparameter for data with KNN imputation

Hyperparameter	Value
Max depth	1
Estimators	10
Min child weight	1
Subsample	1
Colsample by tree	1
Colsample by node	1
Colsample by level	1



(a) Model performance for data with missing value removal

(b) Model performance for data with KNN imputation

(c) Model performance for data without imputation

Fig. 8 Model performance due to the number of variable

5. 결 론

본 연구는 결측치 대체 방법(결측치 제거, KNN 대체)에 따른 차원축소율 차이, PCA방식과 모델 기반의 피처 선택 방식 간의 차원축소율 비교, 그리고 차원축소 방식에 따른 모델 예측성능을 분석하였다. 연구결과, 결측치를 제거한 뒤 PCA를 적용한 경우 KNN 대체보다 원데이터 분산의 95%를 설명하는데 필요한 주성분 개수가 적었다. 또한 PCA를 통해 차원을 축소한 것 보다 특징 선택을 통한 차원 축소가 차원축소율 및 모델의 예측 정

확도 측면에서 더 우수한 결과를 보였다. 흥미로운 점은, 결측치를 전혀 처리하지 않은 XGBoost 모델이 가장 높은 예측 정확도를 나타냈다는 점이다. 이 결과는 XGBoost의 결측치 처리 알고리즘이 일반적인 결측치 처리 방식보다 더 우월할 수 있음을 시사하며, 데이터 분석 과정에서 데이터 전처리 특히 결측치 처리에 들어가는 노력과 비용을 고려할 때 주목할 만한 지점이다.

물론, 본 연구 결과를 일반화하는 것은 한계가 있다. 사용된 데이터가 기상정보 하나뿐이며, 모델 간의 정확도 차이 또한 크지 않기 때문이다. 하지만 결측치를 처리하지 않고도 효과적인 성능을 발휘할 수 있는 모델 개발의 가능성은 중요한 시사점을 제공한다. 이러한 연구 결과는 건물 에너지 데이터 분석에서 결측치 처리 방법의 선택과, XGBoost와 같은 내부적으로 결측치를 처리할 수 있는 모델의 활용을 촉진할 수 있으며 이를 통해 결측치 처리에 드는 비용과 시간을 상당히 감소시킬 수 있다. 향후 연구에서는 더 다양한 유형의 데이터와 더 큰 데이터 셋을 활용하여 본 연구 결과를 검토하고자 한다. 더불어, XGBoost 이외에도 결측치를 내부적으로 처리하는 LightGBM, CatBoost와 같은 다른 앙상블 모델의 성능에 대해서도 평가해 보고자 한다.

후 기

본 연구는 2023년도 산업통상자원부의 재원으로 한국에너지 기술평가원(KETEP) 에너지인력양성사업의 지원을 받아 수행한 연구 과제입니다(No. RS-2023-00237035).

본 연구는 국토교통부/국토교통과학기술진흥원의 지원을 받아 수행한 연구 과제입니다(No. RS-2023-00244769).

REFERENCES

1. Critical Data, M. I. T., Secondary Analysis of Electronic Health Records, Springer Nature, p. 427, 2016.
2. Van Buuren, S., Flexible Imputation of Missing Data, CRC press, 2018.
3. Chen, T. and Guestrin, C., Xgboost: A Scalable Tree Boosting System, In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, August 2016, San Francisco, CA, USA.
4. Mustika, W. F., Murfi, H., and Widyaningsih, Y., Analysis Accuracy of XGboost Model for Multiclass Classification-A Case Study of Applicant Level Risk Prediction for Life Insurance, In 2019 5th International Conference on Science in Information Technology (ICSITech), IEEE, pp. 71-77, October 2019, Yogyakarta, Indonesia.
5. Rusdah, D. A. and Murfi, H. XGBoost in Handling Missing Values for Life Insurance Risk Prediction, SN Applied Sciences, Vol. 2, pp. 1-10, 2020.
6. Korea Meteorological, Open Weather Data Portal, 2023. https://data.kma.go.kr. last accessed on the 12th October 2023.
7. Akande, O., Li, F., and Reiter, J., An Empirical Comparison of Multiple Imputation Methods for Categorical Data, The American Statistician, Vol. 71, No. 2, pp. 162-170, 2017.
8. Zhang, X., Boscardin, W. J., Belin, T. R., Wan, X., He, Y., and Zhang, K., A Bayesian Method for Analyzing

Combinations of Continuous, Ordinal, and Nominal Categorical Data with Missing Values, Journal of Multivariate Analysis, Vol. 135, pp. 43-58, 2015.

9. Ma, Z. and Chen, G., Bayesian Methods for Dealing with Missing Data Problems, Journal of the Korean Statistical Society, Vol. 47, pp. 297-313, 2018.
10. Dewi, K. C., Mustika, W. F., and Murfi, H., Ensemble Learning for Predicting Mortality Rates Affected by Air Quality, In Journal of Physics: Conference Series, IOP Publishing, Vol. 1192, No. 1, 012021, March 2019, Bandung, Indonesia.
11. Kim, T., Ko, W., and Kim, J., Analysis and Impact Evaluation of Missing Data Imputation in Day-Ahead PV Generation Forecasting, Applied Sciences, Vol. 9, No. 1, 204, 2019.
12. Zhang, S., Nearest Neighbor Selection for Iteratively kNN Imputation, Journal of Systems and Software, Vol. 85, No. 11, pp. 2541-2552, 2012.
13. Benesty, J., Chen, J., Huang, Y., and Cohen, I., Pearson Correlation Coefficient, Noise Reduction in Speech Processing, Springer Topics in Signal Processing, Vol. 2, pp. 1-4, 2009.
14. Abdi, H. and Williams, L. J., Principal Component Analysis, Wiley Interdisciplinary Reviews: Computational Statistics, Vol. 2, No. 4, pp. 433-459, 2010.
15. Lee, K. and Lim, H., Correlation Analysis of Building Parameters According to ASHRAE Standard 90.1, Journal of Building Engineering, Vol. 82, 108130, 2023, https://doi.org/10.1016/j.jobe.2023.108130.
16. Chu, H.-G., Shin, H.-S., Cho, S.-K., Yoo, Y.-S., and Park, C.-S., Sensitivity Analysis Using Explainable AI for Building Energy Use, Journal of the Architectural Institute of Korea, Vol. 38, No. 11, pp. 279-287, 2022.
17. Lundberg, S. M. and Lee, S. I., A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017.
18. Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B., Algorithms for Hyper-Parameter Optimization, Advances in Neural Information Processing Systems, Vol. 24, Curran Associates, Inc., 2011.
19. ASHRAE, ASHRAE Handbook: Fundamentals 2021 SI, ASHRAE, 2021.