

머신러닝 기반 전기 에너지 소비량 예측

윤성연 · 오수민 · 김지연 · 박민서[†]

서울여자대학교 데이터사이언스학과

Machine Learning-based Electrical Energy Consumption Prediction

Sungyeon Yoon · Sumin Oh · Jiyun Kim · Minseo Park

Department of Data Science, Seoul Women's University

Electrical energy is a fundamental resource widely used in both daily life and various industrial sectors. Accurate prediction of electricity consumption is critical for managing electrical energy. In this paper, we propose a machine learning model to predict electricity consumption with data from the City of Chicago in Illinois. We modeled and validated five models: Linear Regression, Decision Tree Regression, Random Forest Regression, XGBoost Regression(eXtreme Gradient Boosting Regression), and LightGBM Regression(Light Gradient Boosting Machine Regression). The LightGBM Regression has shown the most suitable model for predicting electricity consumption. We also visualized the effects of independent variables on electricity consumption with XAI(eXplainable Artificial Intelligence), SHAP(SHapley Additive exPlanations). We expect that our proposed model helps predicting and managing electrical energy effectively.

Keywords: Electrical Energy Consumption, Machine Learning, LightGBM Regression, XAI(eXplainable Artificial Intelligence), SHAP(SHapley Additive exPlanations)

1. 서론

전기 에너지는 일상생활뿐만 아니라 산업 전반에서 활용되는 주요 에너지 자원 중 하나이다(Massoud and John, 2008). 인공지능(Artificial Intelligence, AI)의 발달과 함께 생성형 AI(Generative Artificial Intelligence) 시대가 열리면서 전 세계적으로 전기 에너지의 소비가 큰 폭으로 증가하고 있다(Douwes *et al.*, 2021). 오픈 AI(Open AI)의 ChatGPT로 질의를 한 번 주고받을 때마다 소비되는 전기 에너지는 약 2.9 와트시(Watt-hour, Wh)로 기존 구글 검색(Googling)의 약 10배 이상의 전기 에너지가 소비된다. 국제 에너지 기구(International Energy Agency, IEA)는 2026년 전 세계 데이터 센터의 연간 전기 에너지 소비량이 1,050 테라와트시(Terawatt-hour, TWh)에 이를 것으로 전망하고 있으며, 이는 2022년 전 세계 데이터 센터의 연간 전기 에너지 소비량의 약 2.3배

다(International Energy Agency, 2024).

전기 에너지는 생산된 후에는 저장할 수 없고, 또 부족하게 생산할 경우 에너지의 부족으로 인해 큰 경제적 손실을 초래할 수 있다(Ou *et al.*, 2015). 이에, 효율적인 전기 에너지 생산 및 공급량 관리를 위해 전기 에너지 소비량을 예측하는 연구가 활발하게 진행되고 있다(Wang *et al.*, 2018; Hosseini and Fard, 2021; Albuquerque *et al.*, 2022). 그러나 대부분의 선행 연구는 특정 건물의 전기 에너지 소비량만을 예측하거나, 국가 전체의 전기 에너지 소비량을 통합하여 예측하는 데 초점을 맞추고 있다.

전 세계 전기 에너지 소비량의 약 21%를 차지하는 미국의 전기 에너지 시장은 주(State)마다 규제 조건이 상이하다(<https://yearbook.enerdata.co.kr/>). 특히 일리노이 주(State of Illinois), 텍사스 주(State of Texas), 캘리포니아 주(State of California) 등은 전기 에너지 시장이 개방되어 있어 소비자가 자유롭게 다양한 전력 공

[†] 연락저자 : 박민서 교수, 01797 서울특별시 노원구 화랑로 621 서울여자대학교 데이터사이언스학과, Tel: 02-970-7504,

E-mail: mpark@swu.ac.kr

2024년 9월 6일 접수; 2024년 8월 18일; 2024년 11월 25일 수정본 접수; 2024년 11월 28일 게재 확정.

기업체의 전력망을 선택할 수 있다(U.S. Department of Energy, 2015). 이에 따라 해당 주의 전력 회사들은 효율적인 에너지 생산량 관리 전략을 수립하고, 고객 수요에 적합한 판매가격을 책정하기 위해 시(City)의 단위로 그 소비량을 분석한다(U.S. Energy Information Administration, 2018). 전기 에너지 소비량의 정확한 분석을 위해서는 주마다 다른 기후나 공휴일, 업무 시간 등을 함께 고려할 필요가 있다(<https://www.usa.gov/holidays>).

이에 본 논문에서는 미국 중서부의 가장 큰 도시로, 동부와 서부의 전력망을 연결하는 주요 허브인 일리노이 주의 시카고 시(<https://ilenviro.org/energy/>)에 전기 에너지를 공급하고 있는 커먼웰스 에디슨(Commonwealth Edison, ComED)의 시간대별(hourly) 전기 에너지 소비량을 우선 예측(Prediction)하고자 한다. 전기 에너지 소비량의 변화가 시간의 흐름과 관련이 있으나, 그 변화가 반드시 시간적인 패턴만으로 설명되는 것은 아니기 때문이다(Xie *et al.*, 2024). 전기 에너지 소비량 예측은 값에 대한 예측이므로 머신러닝 중 회귀(Regression) 모델을 우선 적용하고자 한다.

회귀 모델은 종속변수와 독립변수 간의 인과관계를 학습하고, 이를 기반으로 새로운 독립변수에 대한 종속변수의 값을 예측하는 모델이다(Zailan *et al.*, 2020; Oh and Park, 2024). 회귀 모델을 사용하여 전기 에너지 소비량 예측과 함께 주요 원인을 찾고자 한다. 5가지 회귀 모델(선형 회귀(Linear Regression), 의사결정 트리 회귀(Decision Tree Regression), 랜덤 포레스트 회귀(Random Forest Regression), XGBoost 회귀(eXtreme Gradient Boosting Regression), LightGBM 회귀(Light Gradient Boosting Machine Regression))로 모델링 및 테스트한 후, 최적의 모델을 제안한다. 또한, 설명 가능한 인공지능(explainable Artificial Intelligence, XAI) 기법 중 하나인 SHAP(SHapley Additional exPlanations)(Salih *et al.*, 2024)을 활용하여 전기 에너지 소비량에 영향을 주는 요인을 시각화한다.

본 논문은 다음과 같이 구성된다. 제2장에서는 전기 에너지 소비량 예측에 관한 선행 연구를 살펴본다. 제3장에서는 회귀 모델에 대해 설명하고, 제4장에서는 머신러닝 기반의 전기 에너지 소비량 예측 모델을 제안 및 결과를 분석한다. 제5장에서는 본 논문의 결론을 언급한다.

2. 전기 에너지 소비량 예측의 선행 연구

Wang *et al.*(2018)은 7개의 날씨 변수, 3개의 시간 변수, 1개의 인구 변수 등 총 11개의 변수를 활용하여 플로리다 대학교(University of Florida)의 시간대별(hourly) 전기 에너지 소비량을 예측하는 머신러닝(Machine Learning) 모델을 제안하였다. 3가지 머신러닝 모델(의사결정 트리 회귀(Decision Tree Regression), 랜덤 포레스트 회귀(Random Forest Regression), SVM 회귀(Support Vector Machine Regression))를 활용하여 모델링(Modeling)하고, 테스트를 통해 랜덤 포레스트 회귀가 가

장 우수한 성능을 보임을 증명하였다. 또한, 실험을 통해 ‘요일(Day Type)’, ‘시(Hour)’, ‘습도(Humidity)’가 전기 에너지 소비량에 큰 영향을 미치는 것을 확인하였다.

Hosseini and Fard(2021)는 8개의 건물 관련 변수를 활용하여 768개 건물의 전기 에너지 소비량을 예측하는 머신러닝 모델을 제안하였다. 3가지 머신러닝 모델(의사결정 트리 회귀, 덤포레스트 회귀, K-최근접 이웃 회귀(K-Nearest Neighbor Regression))를 활용하여 모델링하고, 테스트를 통해 랜덤 포레스트 회귀가 가장 우수한 성능을 보임을 증명하였다. 또한, 실험을 통해 ‘전체 높이(Overall Height)’, ‘지붕 면적(Roof Area)’, ‘표면 면적(Surface Area)’이 건물의 전기 에너지 소비량에 큰 영향을 미치는 것을 확인하였다.

Albuquerque *et al.*(2022)은 185개의 전력 소비량(Power electricity consumption, PEC) 변수, 6개의 날씨 변수, 180개의 날씨 변수, 60개의 전기 에너지 가격 변수, 79개의 경제 변수 등 총 510개의 변수를 활용하여 브라질의 전력 소비량 예측을 위한 통계기법과 머신러닝 모델을 함께 고려하는 방법을 제안하였다. 2개의 통계 기법(자기 회귀 누적 이동평균 모델(Autoregressive Integrated Moving Average Model, ARIMA), 랜덤 워크(Random Walk))과 6개의 머신러닝 모델(Lasso 회귀(Least Absolute Shrinkage and Selection Operator Regression), LARS(Least Angle Regression), Lasso LARS, Ridge 회귀(Ridge Regression), Elastic-Net 회귀(Elastic-Net Regression), 랜덤 포레스트 회귀)를 활용해 모델링하고, 테스트를 통해 랜덤 포레스트 회귀가 가장 우수한 성능을 보임을 증명하였다.

이처럼 선행 연구들은 특정 건물의 전기 에너지 소비량만을 예측하거나, 국가 전체의 전기 에너지 소비량을 통합하여 예측하는 한계가 있다. 그러나 미국은 각 주(State)마다 지정된 공휴일과 업무시간이 상이하기 때문에 이를 고려한 새로운 연구가 필요하다. 따라서, 본 논문에서는 미국 일리노이 주(State of Illinois)의 시카고 시(City of Chicago)의 전기 에너지 소비량을 예측하고자 한다. 시카고 시(City of Chicago)에 전기를 공급하는 커먼웰스 에디슨(Commonwealth Edison, ComED)의 전기 에너지 소비량을 예측하는 머신러닝 모델을 제안한다.

3. 방법론

전기 에너지 소비량 예측은 값에 대한 예측이므로, 머신러닝(Machine Learning) 중 회귀(Regression) 모델을 우선 적용해 보고자 한다. 본 장에서는 값을 예측하는 대표적인 회귀모델 5가지를 살펴 본다: 선형 회귀(Linear Regression), 의사결정 트리 회귀(Decision Tree Regression), 랜덤 포레스트 회귀(Random Forest Regression), XGBoost 회귀(eXtreme Gradient Boosting), LightGBM 회귀(Light Gradient Boosting Machine Regression). 또한, 모델의 결과 도출 과정을 해석할 수 있는 대표적인 설명 가능한 인공지능(explainable Artificial Intelligence, XAI) 기법

인 SHAP(SHapley Additional exPlanations)에 대해 살펴본다.

3.1 연속적인 값 예측에 효과적인 회귀 모델

(1) 선형 회귀(Linear Regression)

선형 회귀는 종속변수와 한 개 이상의 독립변수 간의 관계를 학습하여 종속변수를 설명하거나 예측하는 머신러닝 모델이다(Su *et al.*, 2012). 선형 회귀식(Linear Regression Equation)을 통해 독립변수가 종속변수에 미치는 영향을 파악하고, 가중치(Weight)를 통해 각 독립변수의 중요도를 설명하기 용이하다. 연속적인 값을 가진 독립변수 간의 관계를 직관적으로 표현하는 데에 효과적이다.

(2) 의사결정 트리 회귀(Decision Tree Regression)

의사결정 트리 회귀는 노드(Node)의 의사결정 규칙을 기반으로 데이터를 자식 노드(Child Node)로 분할하는 과정을 반복하고, 마지막 리프 노드(Leaf Node)의 결과를 기반으로 종속변수의 값을 예측하는 머신러닝 모델이다(Navada *et al.*, 2012). 노드의 분할은 자식 노드의 불순도(Impurity)를 기준으로 하며, 이는 평균 제곱 오차(Mean Squared Error, MSE)로 측정한다. 리프 노드에 속하는 데이터의 종속변수의 평균값을 최종 예측값으로 사용한다. 의사결정 트리 회귀 나무 모양으로 시각화할 수 있어 다른 모델에 비해 분석과정을 이해하기 용이하다. 독립변수에 카테고리형 변수가 포함된 경우, 선형 회귀보다 우수한 성능을 보일 수 있다.

(3) 랜덤 포레스트 회귀(Random Forest Regression)

랜덤 포레스트 회귀는 의사결정 트리의 앙상블(Ensemble) 기법 중 하나로 여러 개의 의사결정 트리를 조합하여 종속변수의 값을 예측하는 머신러닝 모델이다(Segal, 2004). 훈련 데이터 셋(Training Sets)의 중복을 허용하여 샘플링(Sampling)하는 배깅(Bootstrap Aggregating, Bagging) 방식을 활용해 각 트리를 학습한다. 트리의 다양성을 위해 독립변수 또한 무작위로 샘플링하여 사용한다. 각 트리의 결과의 평균값을 최종 예측값으로 활용한다. 랜덤 포레스트 회귀는 샘플링된 의사결정 트리를 여러 개 생성하고 조합하기 때문에 과적합(Overfitting)을 방지할 수 있다.

(4) XGBoost 회귀(eXtreme Gradient Boosting Regression)

XGBoost 회귀는 여러 개의 의사결정 트리를 순차적으로 학습하여 종속변수의 값을 예측하는 앙상블 모델이다(Chen and Guestrin, 2016). 이전 트리의 오차(Residual Error)를 기반으로 변수의 가중치(Weight)를 조정하여 모델의 예측 정확도를 점진적으로 개선하는 그라디언트 부스팅(Gradient Boosting) 방식을 사용한다. 부스팅은 예측 결과의 오차를 줄이는 방향으로 진행되며 이는 평균 제곱 오차로 측정한다. 초기 예측값에 모든 부스팅 결과를 조합한 예측 결과를 최종 예측값으로 활

용한다. 균형 트리(Level-wise) 분할 방식을 사용하기 때문에 대규모 데이터에서도 안정적인 학습이 가능하다(Zhang and Gong, 2020).

(5) LightGBM 회귀(Light Gradient Boosting Machine Regression)

LightGBM 회귀는 XGBoost 회귀와 마찬가지로 그라디언트 부스팅 방식을 사용하는 앙상블 모델이다(Ke *et al.*, 2017). 그러나 오차가 큰 리프 노드를 우선적으로 분할하는 리프 중심(Leaf-wise) 분할 방식을 사용하여 다른 모델에 비해 학습시간이 빠르다는 장점이 있다.

3.2 설명 가능한 인공지능(eXplainable Artificial Intelligence, XAI): SHAP(SHapley Additional exPlanations)

설명 가능한 인공지능은 모델이 도출한 과정 및 결과를 사용자가 쉽게 해석할 수 있도록 돕는 기법이다(Gilpin *et al.*, 2018). SHAP은 대표적인 설명 가능한 인공지능 기법으로, 독립변수의 중요도(Feature Importance)를 시각화하여 모델의 결과 도출의 주요 요인을 직관적으로 파악하는데 효과적이다(Fryer *et al.*, 2021). SHAP은 샤플리 값(Shapley Value)을 기반으로 각 독립변수가 모델의 결과에 미치는 영향력을 평가한다(Lundberg and Lee, 2017). 샤플리 값이 양수일 경우 해당 독립변수가 예측값을 증가시키는데, 음수일 경우 예측값을 감소시키는 데 기여한 것으로 해석할 수 있다.

4. 시카고 시의 전기 에너지 소비량 예측 모델 개발

본 연구에서는 시카고 시의 전기 에너지 소비량을 예측하기 위한 머신러닝(Machine Learning) 기반 회귀(Regression) 모델을 설계한다: 데이터 수집(Data Collection), 데이터 전처리(Data Preprocessing), 모델링(Modeling), 검증(Validation)의 네 단계로 구성된다. 제안 연구의 전체 프로세스는 <Figure 1>과 같다. <Figure 1> (a)는 데이터 수집을, <Figure 1> (b)는 데이터 전처리를, <Figure 1> (c)는 머신러닝 기반의 회귀 모델을 활용한 모델링을, <Figure 1> (d)는 모델링의 검증을 의미한다.

4.1 데이터 수집(Data Collection)

연구에 사용한 데이터는 캐글(Kaggle)에서 수집한다(David, 2017; Rob, 2018). 미국 일리노이 주(State of Illinois)의 시카고 시(City of Chicago)에 전기 에너지를 공급하는 커먼웰스 에디슨(Commonwealth Edison, ComED) 회사에서 제공하는 2012년 10월 01일 12시부터 2017년 11월 30일 00시까지의 시간대별(hourly) 전기 에너지 소비량 45,250개를 활용한다. 전기 에너지 소비량은 날씨 변화에 따라 큰 차이를 보이기 때문에 시카고 시의

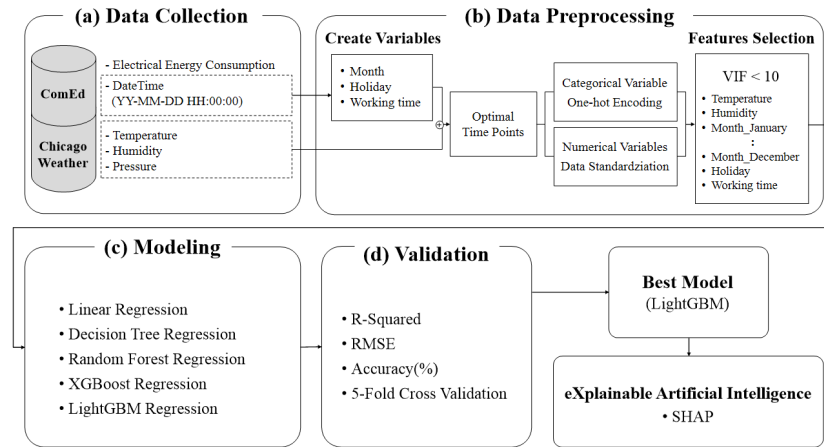


Figure 1. Flowchart of Electrical Energy Consumption Prediction Model

시간대별 기온, 상대습도, 기압 데이터를 함께 활용한다(Argonne National Laboratory, 2022). 전기 에너지 소비량은 메가와트 (Megawatt, MW), 기온은 켈빈(Kelvin, K), 상대습도는 퍼센트 (%), 기압은 표준대기압 1,013 헥토파스칼(Hectopascal, hPa)을 기준 단위로 한다.

모델링을 위한 전체 데이터는 종속변수 열인 ‘전기 에너지 소비량(Electrical Energy Consumption)’과, ‘시각(DateTime)’, ‘기온(Temperature)’, ‘상대습도(Humidity)’, ‘기압(Pressure)’ 등 4개의 독립변수 열, 각 변수에 해당하는 45,250개의 값이 포함된 행으로 구성된다.

4.2 데이터 전처리(Data Preprocessing)

수집한 데이터 중 변수의 일부가 수집되지 않은 1,165개의 행은 결측치로 판단하여 제거한다. ‘년-월-일 시:분:초 (YY-MM-DD HH:00:00)’의 형태로 표시된 ‘시각(DateTime)’을 기준으로, 01월부터 12월까지의 월을 나타내는 독립변수 ‘월 (Month)’을 파생변수로 생성 및 활용한다. 이때, 각 월을 1에서 12의 값으로 사용할 경우 최대 12배의 가중치를 가질 수 있으므로, 원-핫 인코딩(One-hot Encoding)을 활용해 12개의 카테고리(Category)로 구분한다. 주말과 미국의 공휴일을 기준으로 휴일 여부를 나타내는 카테고리형 변수 ‘휴일(Holiday)’을 파생변수로 사용한다. 공공기관의 업무시간 동안 전기 에너지가 평균 약 1438.38 메가와트 더 소비되므로(Appendix A 참조), 시카고 시(<https://www.chicago.gov>)에서 고시하고 있는 요일별 평균 업무시간을 기준으로 업무시간 여부를 나타내는 카테고리형 변수 ‘업무시간(Working Time)’을 파생변수로 사용한다. 시카고 시의 요일별 평균 업무시간은 월요일, 화요일, 목요일 9시부터 17시 30분까지, 수요일 9시부터 17시까지, 금요일 9시부터 15시 30분까지이다. 전체 데이터가 시간대별로 수집되었기 때문에, 30분 단위의 업무시간의 경우 업무 종료 후의 30분을 포함한다. ‘기온(Temperature)’, ‘상대습도(Humidity)’, ‘기압(Pressure)’ 등 날씨와 관련된 독립변수는 측정과 수집에 시차

가 존재하므로, 기준시간보다 4일 전의 값을 사용한다.

서로 다른 단위나 스케일(Scale)을 가진 여러 독립변수를 함께 모델에 활용하는 경우, 큰 값으로 이루어진 독립변수에 편향(Bias)된 모델이 도출될 수 있다. 이에, 서로 다른 분포를 가지는 수치형 독립변수의 값을 평균 0, 표준편차 1의 정규분포 형태로 표준화(Data Standardization)해 그 차이를 감소시킨다(Yoon and Park, 2023).

표준화 한 모든 독립변수는 분산팽창지수(Variance Inflation Factor, VIF)를 활용해 다중공선성 기준 유의 여부를 평가한다(Vittinghoff *et al.*, 2012). 분산팽창지수의 값이 10 이상인 경우 해당 변수가 다른 변수와 높은 상관관계를 가진다고 판단한다. <Table 1>은 분산팽창지수 기준 유의한 독립변수이다. ‘시각(DateTime)’, ‘기압(Pressure)’의 경우, 분산팽창지수 기준 유의하지 않아 독립변수에서 제외한다.

Table 1. Variance Inflation Factors of Independent Variables

Independent Variables	Variance Inflation Factor
Temperature	3.78
Humidity	1.17
Month_January	1.53
Month_February	1.50
Month_March	1.27
Month_April	1.13
Month_May	1.16
Month_June	1.35
Month_July	1.52
Month_August	1.50
Month_September	1.35
Month_October	1.16
Month_November	1.17
Month_December	1.30
Working Time	1.20
Holiday	1.18

4.3 모델링(Modeling)

데이터 수집과 전처리 후 도출된 16개의 변수로 구성된 44,085개의 데이터를 선형 회귀(Linear Regression), 의사결정 트리 회귀(Decision Tree Regression), 랜덤 포레스트 회귀(Random Forest Regression), XGBoost 회귀(eXtreme Gradient Boosting Regression), LightGBM 회귀(Light Gradient Boosting Machine Regression) 모델에 적용하여 학습한다. 모델링에 사용한 각 모델의 하이퍼파라미터(Hyper-parameter)는 <Table 2>와 같다.

실험을 위해 전체 데이터에서 훈련 데이터(Training Sets)와 테스트 데이터(Test Sets)를 무작위로 각각 80%, 20%의 비율로 나누어 구성한다. 전기 에너지 소비량 예측 모델의 성능을 검증하기 위해 5-겹 교차 검증(5-Fold Cross Validation)을 수행한다.

4.4 검증(Validation)

본 연구에서는 전기 에너지 소비량 예측을 위한 최적의 모

델을 선정하기 위해 정량적 평가와 정성적 평가를 수행하였다. 전기 에너지 소비량 예측 모델을 정량적으로 평가하기 위해 평균 제곱근 오차(Root Mean Squared Error, RMSE), 결정 계수(R-Squared), 정확도(Accuracy)를 측정하였다(Appendix B 참조). 평균 제곱근 오차는 모델의 예측력을 나타내는 지표로, 실제값과 모델이 예측한 값 간의 차이를 기준으로 산출한다. 평균 제곱근 오차의 값이 작을수록 모델의 예측력이 높다고 해석할 수 있다. 결정 계수는 종속변수에 대한 독립변수의 설명력을 정량적으로 평가하기 위한 지표로, 그 값이 1에 가까울수록 모델의 설명력이 우수하다고 해석할 수 있다. 정확도는 모델의 예측 성능을 평가하는 비율로, 그 비율이 100%에 가까울수록 우수한 모델이라고 해석할 수 있다(Lee *et al.*, 2024).

<Table 3>은 5개의 회귀 모델(선형 회귀(Linear Regression), 의사결정 트리 회귀(Decision Tree Regression), 랜덤 포레스트 회귀(Random Forest Regression), XGBoost 회귀(eXtreme Gradient Boosting Regression), LightGBM 회귀(Light Gradient Boosting Machine Regression))의 정량적 성능 평가 결과이다.

Table 2. Hyper-parameter of Electrical Energy Consumption Prediction Models

Model	Number of Trees	Max Depth	Criterion	Cost Function	Max Features	Max Samples	Learning Rate
Linear Regression	N/A	N/A	MSE	N/A	N/A	N/A	N/A
Decision Tree Regression	1	5	MSE	N/A	N/A	N/A	N/A
Random Forest Regression	100	5	MSE	N/A	12(75%)	26,451(75%)	N/A
XGBoost Regression (eXtreme Gradient Boosting Regression)	100	5	MSE	MSE	N/A	N/A	0.1
LightGBM Regression (Light Gradient Boosting Machine Regression)	100	5	MSE	MSE	N/A	N/A	0.1

Table 3. Quantitative Evaluation of Electrical Energy Consumption Prediction Models

Model	Datasets	Evaluation		
		RMSE	R-Squared	Accuracy
Linear Regression	Training Sets	1706.31	0.3934	62.72%
	5-Fold Cross Validation	1712.24	0.3914	62.56%
	Test Sets	1731.09	0.3851	62.05%
Decision Tree Regression	Training Sets	1345.40	0.6229	78.92%
	5-Fold Cross Validation	1348.55	0.6225	78.90%
	Test Sets	1358.57	0.6212	78.82%
Random Forest Regression	Training Sets	1301.83	0.6469	80.43%
	5-Fold Cross Validation	1305.01	0.6465	80.40%
	Test Sets	1315.13	0.6451	80.31%
XGBoost Regression (eXtreme Gradient Boosting Regression)	Training Sets	1108.99	0.7438	86.24%
	5-Fold Cross Validation	1124.91	0.7373	85.87%
	Test Sets	1174.43	0.7170	84.67%
LightGBM Regression (Light Gradient Boosting Machine Regression)	Training Sets	1160.93	0.7192	84.80%
	5-Fold Cross Validation	1168.23	0.7167	84.66%
	Test Sets	1191.27	0.7088	84.19%

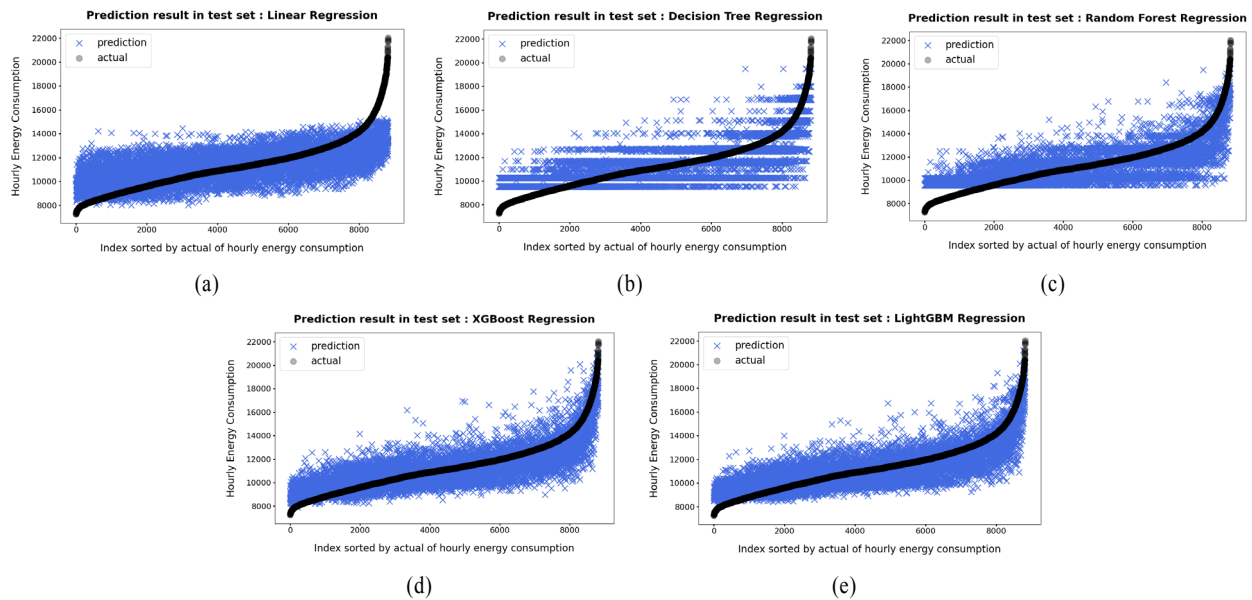


Figure 2. Qualitative evaluation graph of electrical energy consumption prediction models: (a) Linear Regression, (b) Decision Tree Regression, (c) Random Forest Regression, (d) XGBoost Regression(eXtreme Gradient Boosting Regression), (e) LightGBM Regression(Light Gradient Boosting Machine Regression), Black ‘○’ means predicted value and blue ‘×’ means actual value

선형 회귀의 경우, 모든 성능 지표에서 가장 낮은 성능을 보였다. 이는 전기 에너지 소비량을 예측하는 데에 사용한 16개의 독립변수 중 14개가 카테고리형 변수이기 때문이다. 연속적인 값을 가진 독립변수 간의 관계를 직관적으로 표현하는 데에 효과적인 선형 회귀는 전기 에너지 소비량 예측에 적합하지 않은 것을 확인하였다. 의사결정 트리 회귀의 경우, 선형 회귀보다 테스트 데이터(Test Sets)의 정확도에서 약 27.03% 향상된 성능을 보였다. 이는 의사결정 트리가 카테고리형 변수를 다루는 데에 효과적이기 때문이다. 오버피팅(Overfitting)을 줄이기 위해 설계된 앙상블(Ensemble) 모델인 랜덤 포레스트 회귀는 의사결정 트리 모델보다 테스트 데이터의 정확도에서 약 1.89% 향상된 성능을 보였다. 중요도를 함께 고려한 앙상블 모델인 XGBoost 회귀와 LightGBM 회귀의 테스트 데이터 정확도는 각각 84.67%, 84.19%로 랜덤 포레스트 회귀의 테스트 정확도보다 약 5.43%, 4.83% 향상된 우수한 성능을 보였다. XGBoost 회귀와 LightGBM 회귀의 5-겹 교차 검증(5-Fold Cross Validation) 결과 역시 안정적인 성능을 보였다.

전기 에너지 소비량 예측 모델을 정성적으로 평가하기 위해 <Figure 2>와 같이 무작위로 선택한 테스트 데이터(Test Sets)의 실제값과 예측값을 시각화하여 분석하였다. <Figure 2> (a)는 선형 회귀, <Figure 2> (b)는 의사결정 트리 회귀, <Figure 2> (c)는 랜덤 포레스트 회귀, <Figure 2> (d)는 XGBoost 회귀, <Figure 2> (e)는 LightGBM 회귀의 그래프이다. 각 그래프의 x축은 테스트 데이터를 실제값을 기준으로 오름차순한 뒤 새롭게 부여한 인덱스(Index)를, y축은 인덱스에 해당하는 전기 에너지 소비량을 나타낸다. 그래프의 검은색 ‘○’는 실제값을, 파란색 ‘×’는 예측값을 나타낸다. 실제값 ‘○’와 예측값 ‘×’

간의 차이가 작을수록, 전기 에너지 소비량을 정확하게 예측한 것으로 해석할 수 있다. 선형 회귀(<Figure 2> (a))는 16,000 메가와트(Megawatt, MW) 이상의 전기 에너지 소비량을 전혀 예측하지 못하는 모습을 확인할 수 있었다. 의사결정 트리 회귀(<Figure 2> (b))는 구간별로 전기 에너지 소비량을 같거나 유사한 값으로 예측하는 모습을 확인할 수 있었다. 여러 의사결정 트리의 앙상블인 랜덤 포레스트 회귀(<Figure 2> (c))도 구간의 간격을 감소하였으나, 여전히 구간별로 전기 에너지 소비량을 유사한 값으로 예측하는 모습을 확인할 수 있었다. XGBoost 회귀(<Figure 2> (d))와 LightGBM 회귀(<Figure 2> (e))는 전기 에너지 소비량을 유사한 추세로 우수하게 예측하나, XGBoost 회귀(<Figure 2> (d))가 8,000 메가와트에서 8,500 메가와트 사이의 전기 에너지 소비량을 좀 더 정확하게 예측하는 것을 확인할 수 있었다. 본 연구에서는 전기 에너지 소비량을 예측하기 위한 최적의 모델로 LightGBM 회귀를 제안하였다. 정량적 평가에서 가장 높은 성능을 보였던 XGBoost 회귀보다 테스트 데이터의 정확도 기준 약 0.57%의 성능 하락을 보이나, 훈련 데이터(Training Sets)와 테스트 데이터 간의 성능 차이가 61.15% 감소하였기 때문이다. 이는 LightGBM 회귀 모델이 오버피팅(Over-Fitting) 문제를 XGBoost 회귀보다 더 잘 해결했다고 해석할 수 있다. 또한 <Table 4>와 같이 LightGBM 회귀 모델의 리프 중심(Leaf-wise) 분할 방식은 모델의 학습 시간과 메모리 사용량을 감소시킨다는 장점이 있다. 이러한 특성은 추후 대규모 데이터셋으로 확장하여 분석할 때 더욱 효과적이다. 대규모 데이터셋을 처리하기 위해서는 시간적, 물리적 비용의 효율화가 필수적이기 때문이다 (Ke et al., 2017).

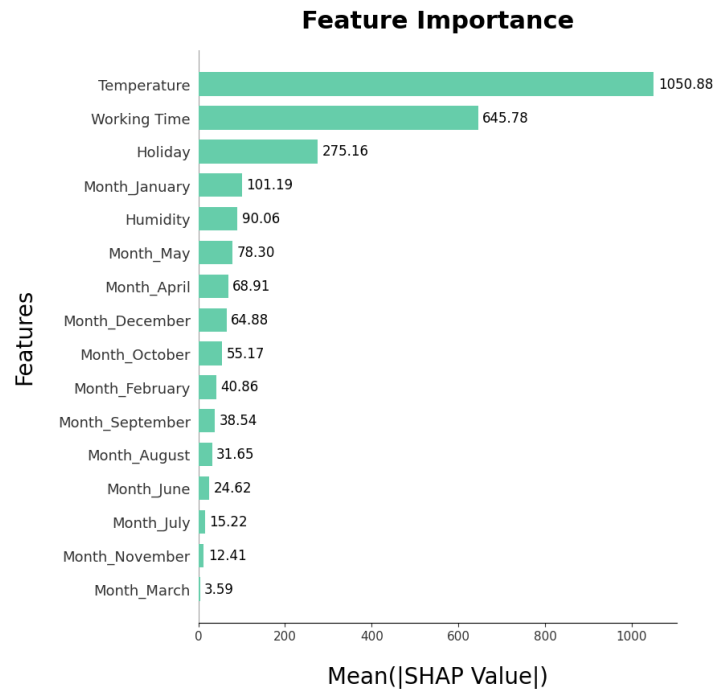
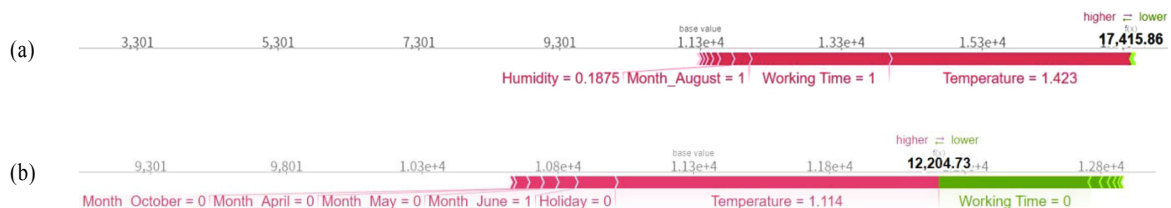
Table 4. Evaluation of Training Times and Memories for XGBoost Regression and LightGBM Regression

Model	Times(Seconds)	Memory(MB)
XGBoost Regression (eXtreme Gradient Boosting Regression)	1.0819	412.27
LightGBM Regression (Light Gradient Boosting Machine Regression)	0.2942	359.23

모델의 결과 도출 과정을 직관적으로 이해하고 주요 원인을 파악하는데 효과적인 설명 가능한 인공지능(eXplainable Artificial Intelligence, XAI) 기법으로 한 번 더 검증하였다. <Figure 3>은 설명 가능한 인공지능의 대표적인 기법인

SHAP(SHapley Additional exPlanations)으로 시각화한 결과이다. 제안 모델(LightGBM 회귀)의 독립변수의 중요도(Feature Importance)를 계산한 뒤, 내림차순으로 정렬한 결과이다. ‘기온(Temperature)’, ‘업무시간(Working Time)’, ‘휴일(Holiday)’ 순으로 전기 에너지 소비량에 기여도가 높은 것을 확인하였다. 이를 통해 시카고 시의 전기 에너지 소비량에 영향을 주는 주요 요인을 알 수 있었다.

<Figure 4>는 제안 모델(LightGBM 회귀)의 테스트에 사용한 데이터 중 일부를 SHAP으로 시각화한 결과이다. <Figure 4> (a)는 가장 정확하게 예측한 데이터(2016년 08월 25일 17시의 전기 에너지 소비량)의 결과이고, <Figure 4> (b)는 가장 예측력이 떨어지는 데이터의(2014년 06월 17일 21시의 전기 에너지 소비량) 결과이다. 분홍색은 예측값을 증가시키는 데에 기여하는 독립변수를, 초록색은 예측값을 낮추는 데에 기여하는 독립변수를 의미한다. 가장 정확하게 예측한 데이터(2014

**Figure 3.** LightGBM Regression's Feature Importance Calculated by the SHAP(SHapley Additional exPlanations)**Figure 4.** Key Independent Variables that Affect Electrical Energy Consumption: (a) Best Prediction by LightGBM Regression, (b) Worst Prediction by LightGBM Regression

년 06월 17일 21시의 전기 에너지 소비량)에 대해 제안 모델(LightGBM 회귀)은 전기 에너지 소비량을 17,415.86으로 예측하였고, 실제 값은 17,415였다. 전기 에너지 소비량에 가장 큰 영향을 준 독립변수는 '기온(Temperature)'이었다. 즉, 기온이 높아지면 전기 에너지 소비량이 늘어날 확률이 높아진다고 해석할 수 있다. 가장 예측력이 떨어지는 데이터(2014년 06월 17일 21시의 전기 에너지 소비량)에 대해 제안 모델(LightGBM 회귀)은 전기 에너지 소비량을 12,204.73으로 예측하였고, 실제 값은 18,032였다. 이는 4시간 전 시카고를 포함한 일리노이주 북부지역에 폭풍우로 인한 대규모 정전이 발생했기 때문이었다(Stephens, 2014). 자연재해로 인한 정전 후 전력망이 복구되는 과정에서 전기 에너지 소비량이 급증함을 알 수 있었다.

5. 결 론

본 논문에서는 머신러닝(Machine Learning)을 활용하여 미국 일리노이 주(State of Illinois) 시카고 시(City of Chicago)의 전기 에너지 소비량을 예측하는 방법을 제안하였다. 커먼웰스 에디슨(Commonwealth Edison, ComED)에서 제공하는 시간대별(hourly) 전기 에너지 소비량과 시간대별 날씨 관련 변수를 활용하였다. 변수의 일부가 수집되지 않은 행을 결측치로 판단하여 제거하였다. 01월부터 12월까지의 월을 나타내는 12개의 카테고리형 변수를 파생변수로 생성하였다. 주말과 미국의 공휴일을 기준으로 휴일 여부를 나타내는 카테고리형 변수를 파생변수로 생성하였다. 요일별 평균 업무시간을 기준으로 업무시간 여부를 나타내는 카테고리형 변수를 파생변수로 생성하였다. 데이터의 측정과 수집에 시차가 존재하는 변수에 시차를 반영하였다. 서로 다른 분포를 가지는 수치형 독립변수의 값을 평균 0, 표준편차 1의 정규분포 형태로 표준화(Data Standardization)하였다. 독립변수의 전처리(Data Preprocessing) 후, 분산팽창지수(Variance Inflation Factor, VIF)를 활용하여 다중공선성 기준 유의 여부를 평가하였다. 최종 선정한 16개의 독립변수를 활용해 5개의 머신러닝 기반 회귀(Regression) 모델로 모델링하고 테스트 및 검증하였다: 선형 회귀(Linear Regression), 의사결정 트리 회귀(Decision Tree Regression), 랜덤 포레스트 회귀(Random Forest Regression), XGBoost 회귀(eXtreme Gradient Boosting Regression), LightGBM 회귀(Light Gradient Boosting Machine Regression). 모델의 성능을 검증하기 위해 5-겹 교차 검증(5-Fold Cross Validation)을 수행하였다. 그 결과 시카고 시의 전기 에너지 소비량을 예측하기 위한 모델로 LightGBM 회귀를 최종 제안하였다. 제안 모델의 결과 도출 과정을 직관적으로 이해하고 주요 원인을 명확히 명시하기 위해 SHAP(SHapley Additional exPlanations)을 활용하여 시각화하였다.

본 연구에서 제안한 전기 에너지 소비량 예측 모델은 도출 과정과 결과에 대한 해석력을 가지고 있어 높은 신뢰성과 안

정적인 성능을 보인다는 데에 큰 의의가 있다. 이는 전력 회사의 효율적인 전기 에너지 생산량 관리에 도움이 될 수 있다. 또한, 소비량을 기반으로 고객 수요를 분석해 판매가격을 탄력적으로 책정하는 데에 도움이 될 수 있다. 그러나, 커먼웰스 에디슨의 전기 에너지 소비량 데이터가 시카고 시 이외의 지역의 전기 에너지 소비량이 통합되어 있어, 타 지역의 소비 특성이 일부 반영되었다는 한계가 존재한다. ‘업무시간(Working Time)’ 변수에 비표준 근무가 일반적인 시카고 산업체(Henly *et al.*, 2006)의 업무시간을 반영하지 못했다는 한계가 존재한다. 수집 가능한 데이터가 2012년에서 2017년까지의 데이터로, 생성형 AI(Generative Artificial Intelligence) 시대가 열린 후 증가한 전기 에너지 소비량을 반영하는 데에 한계가 존재한다. 이전 시점의 전기 에너지 소비량을 독립변수로 함께 활용하는 경우 변수의 독립성 가정에 위배될 수 있어(Nimon, 2012; Flatt and Jacobs, 2019) 제안 모델에서는 이를 변수로 포함하지 않았으나, 전기 에너지 소비량의 주기적 특성을 고려하여 해당 변수를 포함할 경우 모델의 성능이 향상되는 것을 확인하였다(<Appendix C> 참조). 향후, 최근의 전기 에너지 소비량 데이터와 지역적 국제적 환경 관련 변수를 추가적으로 고려한 모델을 개발하면 보다 정확한 전기 에너지 소비량 예측이 가능할 것으로 기대한다.

참고문헌

- Albuquerque, P. C., Cajueiro, D. O., and Rossi, M. D. C. (2022), Machine learning models for forecasting power electricity consumption using a high dimensional dataset, *Expert Systems with Applications*, **187**, 1-13.
- Argonne National Laboratory (2022), ComEd Climate Risk and Adaptation Outlook, Phase 1: Temperature, Heat Index, and Average Wind.
- Chen, T. and Guestrin, C. (2016), XGBoost: A Scalable Tree Boosting System, arXiv preprint, arXiv:1603.02754.
- Cox, H. J. and Armstrong, J. H. (1914), The Weather and Climate of Chicago, *Geographic Society of Chicago*, **4**, 1943.
- David, B. (2017), Historical Hourly Weather Data 2012-2017, Kaggle, <https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data>.
- Douwes, C., Esling, P., and Briot, J. P. (2021), Energy Consumption of Deep Generative Audio Models, arXiv preprint, arXiv:2107.02621.
- Flatt, C. and Jacobs, R. L. (2019), Principle assumptions of regression analysis: Testing, techniques, and statistical reporting of imperfect data sets, *Advances in Developing Human Resources*, **21**(4), 484-502.
- Fryer, D., Strumke, I., and Nguyen, H. (2021), Shapley values for feature selection: The good, the bad, and the axioms, arXiv preprint, arXiv:2102.10936.
- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018), Explaining explanations: An overview of interpretability of machine learning, 80-89.
- Henly, J. R., Shaefer, H. L., and Waxman, E. (2006), Nonstandard

- work schedules: Employer-and employee-driven flexibility in retail jobs, *Social Service Review*, **80**(4), 609-634.
- International Energy Agency (2024), Analysis and forecast to 2026.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T. (2017), LightGBM: A Highly Efficient Gradient Boosting Decision Tree, *Advances in Neural Information Processing Systems*, 30.
- Lee, H., Sohn, S., and Park, M. (2024), Deep Learning-Based Daily Baseball Attendance Prediction, *The Journal of the Convergence on Culture Technology*, **10**(3), 131-135.
- Lundberg, S. M. and Lee, S. I. (2017), A Unified Approach to Interpreting Model Predictions, *Neural Information Processing*, 17.
- Massoud, A. and John, S. (2008), The Electric Power Grid: Today and Tomorrow, *Carriers, Storage, & Transformation/Electrical Flow*, **33**(4), 399-407.
- Navada, A., Ansari, A. N., Patil, S., and Sonkamble, B. A. (2012), Overview of use of decision tree algorithms in machine learning, *2011 IEEE Control and System Graduate Research Colloquium*, 37-42.
- Nimon, K. (2012), Statistical assumptions of substantive analyses across the general linear model: A mini review, *Frontiers in Psychology*, **3**, 322.
- Oh, S. and Park, M. (2024), Deep Learning-based Happiness Index Model Considering Social Variables and Individual Emotional Index, *The Journal of the Convergence on Culture Technology*, **10**(1), 489-493.
- Ou, P., Huang, R., and Yao, X. (2016), Economic Impacts of Power Shortage, *Sustainability*, **8**(7), 687.
- Rob, M. (2018), Hourly Energy Consumption, Kaggle, <https://www.kaggle.com/datasets/robikscube/hourly-energy-consumption>.
- Salih, A. M., Estabragh, Z. R., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., and Menegaz, G. (2024), A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME, *Advanced Intelligent Systems*, 2400304.
- Segal, M. R. (2004), Machine learning benchmarks and random forest regression, UCSF: Center for Bioinformatics and Molecular Biostatistics, <https://escholarship.org/uc/item/35x3v9t4>.
- Stephens, G. (2014), Power Out For Thousands in Northwest Illinois After Monday's Storms, WNIJ WNIU, <https://www.northernpublicradio.org/illinois/2014-06-17/power-out-for-thousands-in-northwest-illinois-after-mondays-storms>.
- Su, X., Yan, X., and Tsai, C. (2012), Linear Regression, *Wiley Interdisciplinary Reviews: Computational Statistics*, **4**(3), 275-294.
- U.S. Department of Energy (2015), United States Electricity Industry Primer.
- U.S. Energy Information Administration (2018), A Guide to EIA Electric Power Data.
- Vittinghoff, E., Glidden, D. V., Shiboski, S. C., and McCulloch, C. E. (2012), Predictor Selection, Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models, 140-144.
- Wang, Z., Wang, Y., Zeng, R., Srinivasan, R. S., and Ahrentzen, S. (2018), Random Forest based hourly building energy prediction, *Energy and Buildings*, **171**, 11-25.
- Xie, P., Wang, S., Liao, J., and Sun, F. (2024), Study on Spatial-Temporal Disparities and Factors Influencing Electricity Consumption Carbon Emissions in China, *Sustainability*, **16**(10), 4128.
- Yoon, S. and Park, M. (2023), Prediction of Gasoline Orders at Gas Stations in South Korea Using VAE-Based Machine Learning Model to Address Data Asymmetry, *Applied Sciences*, **13**(20), 11124.
- Zailan, K. A. B. M., Hasan, M. H. B., and Witjaksono, G. (2020), Comparative analysis of machine learning algorithms for optimizing variable step-size least mean square in motion artifact reduction, *An International Journal of Advanced Computer Technology*, **9**(3), 3590-3595.
- Zhang, D. and Gong, Y. (2020), The comparison of LightGBM and XGBoost coupling factor analysis and prediagnosis of acute liver failure, *IEEE Access*, **8**, 220990-221003.

Appendix A

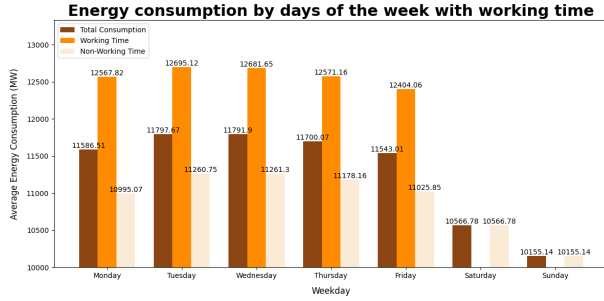


Figure A. Average electrical energy consumption by days of the week with working time: the brown bar means the average hourly electrical energy consumption, the orange bar means the average hourly electrical energy consumption during working time, and the beige bar means the average hourly electrical energy consumption during non-working time.

<Figure A>는 전기 에너지 소비 패턴을 업무시간(Working Time)을 기반으로 분석한 그래프이다. 갈색 막대는 각 요일의 시간대별(hourly) 전기 에너지 소비량의 평균값이다. 주황색 막대와 베이지색 막대는 각각 업무시간과 업무시간 외의 시간에 해당하는 시간대별 전기 에너지 소비량의 평균값을 나타낸다. 전기 에너지는 주말보다 평일에 더 많이 소비되었다. 특히, 평일 업무시간에는 업무시간 외의 시간보다 평균 약 1438.38 메가와트(Megawatt, MW) 더 소비된 것을 알 수 있었다. 가장 큰 차이를 보인 요일은 월요일로, 업무시간 외의 시간보다 약 1568.82 메가와트의 전기 에너지를 더 소비한 것을 알 수 있었다.

Appendix B

$$R_Squared = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \frac{1}{n} \sum_{i=1}^n y_i)^2} \quad (1)$$

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

$$Accuracy = \sqrt{1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \frac{1}{n} \sum_{i=1}^n y_i)^2}} \times 100 \quad (3)$$

식 (1)은 결정 계수(R-Squared)를 계산하는 수식이다. 식 (2)는 평균 제곱근 오차(Root Mean Squared Errors, RMSE)를 계산하는 수식이다. 식 (3)은 정확도(Accuracy)를 계산하는 수식이다. n 은 데이터 세트의 수, y_i 는 i 번째 실제 값, $\hat{y}_i$ 는 i 번째 예측 값을 의미한다.

Appendix C

<Table A>는 제안 모델과 이전 시점 전기 에너지 소비량 변수를 추가한 모델을 비교한 정량적 성능 평가 결과이다. 날씨와 관련된 독립변수에 적용한 4일 전 시차를 동일하게 적용한 ‘이전 시점 전기 에너지 소비량(Previous electric energy consumption)’을 파생변수로 사용하였다. 이전 시점 전기 에너지 소비량 변수를 추가한 모델의 경우, 제안 모델보다 테스트 데이터의 정확도에서 약 4.60% 향상된 성능을 보였다.

<Figure B>는 이전 시점 전기 에너지 소비량 변수를 추가한 모델의 독립변수의 중요도를 계산한 뒤, 내림차순으로 정렬한 결과이다. ‘기온(Temperature)’, ‘이전 시점 전기 에너지 소비량(Previous electric energy consumption)’, ‘휴일(Holiday)’ 순으로 전기 에너지 소비량에 기여도가 높은 것을 확인하였다. 이전 시점 전기 에너지 소비량이 종속변수를 예측하는 데 큰 영향을 미치는 것을 알 수 있었다. 이전 시점 전기 에너지 소비량 변수는 과거 데이터로, 현재 소비량을 예측하는 데 중요한 정보를 제공하기 때문이다. 그러나, 이전 시점 전기 에너지 소비량은 종속변수와 높은 상관성을 가지기 때문에 변수의 독립성 가정에 위배될 수 있다(Nimon, 2012; Flatt and Jacobs, 2019). 이는 종속변수와 직접적인 의존 관계를 형성하여 모델의 설명 가능성과 실제 활용도를 저하시킬 수 있다. 이에 본 연구는 이

Table A. Quantitative Evaluation of the Proposed Model and the Model Including Previous Electrical Energy Consumption

Model	Datasets	Evaluation		
		RMSE	R-Squared	Accuracy
Ours (LightGBM Regression)	Training Sets	1072.60	0.7608	87.22%
	5-Fold Cross Validation	1086.15	0.7553	86.91%
	Test Sets	1128.45	0.7380	85.91%
+ Previous electrical energy consumption (LightGBM Regression)	Training Sets	929.22	0.8201	90.56%
	5-Fold Cross Validation	938.76	0.8171	90.39%
	Test Sets	968.64	0.8075	89.86%

전 시점 전기 에너지 소비량을 제외하여 보다 일반화된 모델을 제안하였다.

by the SHAP(SHapley Additional exPlanations)

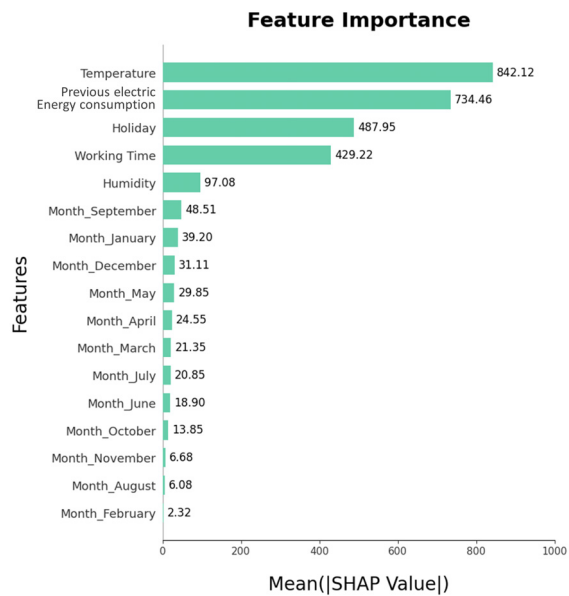


Figure B. LightGBM Regression including previous electrical energy consumption's feature importance calculated

저자소개

윤성연: 서울여자대학교에서 데이터사이언스학과에 재학 중이다. 연구분야는 머신러닝, 딥러닝이다.

오수민: 서울여자대학교에서 데이터사이언스학과에 재학 중이다. 연구분야는 머신러닝, 딥러닝이다.

김지연: 서울여자대학교에서 데이터사이언스학과에 재학 중이다. 연구분야는 머신러닝, 딥러닝이다.

박민서: 2009년 매사추세츠대학교 컴퓨터 사이언스(머신러닝) 전공으로 박사학위를 취득하였다. 삼성SDS Bioinformatics Lab 및 성균관대학교 삼성융합의과학원 수석연구원, SK텔레콤 데이터사업부 팀 리더, 한화시스템 상무(AI 사업담당/AI Lab 장)를 거쳐 현재 KAIST 기술경영학부 겸직교수와 서울여자대학교 데이터사이언스학과 교수로 재직하고 있다. 연구분야는 머신러닝, 딥러닝이다.