과목: 기계학습
과제명: 회귀 예측 모델 구현 및 분석
20 학번 / 이름: 202021065 / 최세진
제출일: 2025.04.26

1. 프로젝트 개요

-문제 정의 :
본 프로젝트에서는 Energy Efficiency 데이터셋을 활용하여 건물의 구조적 특성들을 기반으로 냉방 부하(Cooling Load)를 예측하는 회귀 모델을 구축하였다. 이는 연속형 값을 예측하는 문제로, 다양한 머신러닝 모델을 적용해 성능을 비교하고 최적화하는 것이 목표다.

-데이터셋 설명
1. 출처: Energy Efficiency Dataset - Kaggle
2. 데이터 크기: 768 개 샘플, 8 개 특성
3. 타겟 변수: Cooling Load (Y2)
4. 주요 입력 변수: Relative Compactness, Surface Area, Wall Area, Roof Area, Overall Height, Orientation, Glazing Area, Glazing Area Distribution

2. 데이터 전처리 및 탐색적 분석 (EDA)

-결측치 처리 :
데이터셋에는 결측치가 존재하지 않아 별도의 처리 과정을 수행하지 않았다.

-범주형 변수 처리 :
모든 입력 변수는 수치형 데이터로 되어 있어 별도의 인코딩 과정이 필요하지 않았다.

-스케일링 :
StandardScaler 를 사용하여 입력 특성들을 표준화(평균 0, 표준편차 1)하였다.

-EDA 시각화 :
1. Cooling Load 값의 분포를 히스토그램 및 박스플롯으로 시각화하여 이상치 유무와 데이터의 분포 특성을 분석하였다.
2. 주요 특성들과 Cooling Load 간의 상관관계를 분석하여 어떤 변수들이 예측에 중요한지를 파악하였다.

3. 모델 구축 및 학습

-사용한 알고리즘 :
Linear Regression, RandomForestRegressor

-데이터 분할 방식 :
train_test_split 을 이용하여 학습 데이터와 테스트 데이터를 8:2 비율로 나누었다.

-파이프라인 사용 여부 :
파이프라인은 별도로 구성하지 않고, 전처리와 모델 학습을 개별적으로 수행하였다.

-학습 코드 요약 :
1. StandardScaler 를 통해 입력 데이터를 스케일링
2. 선형 회귀 모델과 랜덤 포레스트 모델을 각각 학습
3. 테스트 데이터에 대해 예측 수행 및 평가

4. 성능 평가

-사용한 지표
1. RMSE (Root Mean Squared Error)
2. MAE (Mean Absolute Error)
3. R^2 Score (결정계수)

-예측값 vs 실제값 시각화
예측 결과와 실제 Cooling Load 값을 산점도로 비교하여 모델의 예측 정확성을 직관적으로 확인하였다.

-해석
1. 선형 회귀 모델은 기본적인 성능을 보였지만, 복잡한 패턴을 충분히 포착하지는 못했다.
2. 반면 RandomForest 모델은 높은 결정계수를 기록하며 실제 값을 효과적으로 예측하였다. 과소/과대 예측 경향은 미미하였다.

5. 하이퍼파라미터 튜닝

-튜닝 방법 :
GridSearchCV 를 활용하여 RandomForest 모델의 최적 하이퍼파라미터를 탐색하였다.

-튜닝한 하이퍼파라미터
1. n_estimators: 100, 200
2. max_depth: 5, 10, None
3. min_samples_split: 2, 5, 10

-튜닝 결과 분석
1. 최적 파라미터 조합은 { 'n_estimators': 200, 'max_depth': 10, 'min_samples_split': 2 }로 확인되었으며,
2. 튜닝을 통해 RMSE 가 감소하고 R^2가 상승하여 모델의 예측력이 전반적으로 향상되었다.

6. 결론 및 고찰

-최종 모델 성능 종합 평가 :
튜닝된 RandomForestRegressor 모델은 매우 우수한 성능을 보여주었고, R^2 Score 는 0.99 에 근접하였다.

-데이터 또는 모델의 한계 :
데이터셋이 깨끗하여 추가적인 결측치나 이상치 처리는 필요 없었지만, 보다 다양한 건물 특성 데이터가 있었다면 일반화 성능이 더욱 높아졌을 것이다.

-실생활 응용 가능성 또는 확장 방향 :
본 모델은 건축 설계 초기 단계에서 에너지 효율을 고려한 건물 설계를 지원하는 데 유용하게 활용될 수 있다.

-다음 단계에서 고려할 점 :
Feature selection 기법 적용
Gradient Boosting 과 같은 다른 앙상블 기법 시도
잔차 분석을 통한 모델 개선

7. 참고자료

-Energy Efficiency Dataset - Kaggle

-Scikit-learn Documentation (v1.3.0)

8. 부록

중요 코드 스니펫

```
# 데이터 분할 및 스케일링

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)


# 모델 학습 및 튜닝
rf = RandomForestRegressor()
rf.fit(X_train_scaled, y_train)


param_grid = {'n_estimators': [100, 200], 'max_depth': [5, 10, None], 'min_samples_split'
: [2, 5, 10]}
grid_search = GridSearchCV(rf, param_grid, cv=5)
grid_search.fit(X_train_scaled, y_train)
```