연료 절감 및 온실가스 감축효과 기대

RSS : 실제 값과 예측 값의 오차로, 간단하고 직관적으로 해석할 수 있다.
MSE : 평균 제곱 오차. 즉 RSS에서 데이터 수만큼 나눈 값이다. 단, outlier에 민감하다.
MAE : 평균 절댓값의 오차. 즉 실제 값과 예측 값 간 오차의 절댓값의 평균이다. 변동성이 큰 지표와 낮은 지표를 같이 예측할 때 유용하다. 또한 가장 간단하여 직관적인데, 다만 평균을 이용하므로 입력 값의 크기에 의존한다.
R^2(결정 계수) : 회귀 모델의 설명력을 표현하는 지표로, 1에 가까울수록 높은 성능을 나타낸다.
ex) 1-(RSS/MSE)

L1 정규화(Lasso) : 불필요한 입력 값에 대응되는 Bi(계수)를 0으로 만든다.
-> Lasso Regression : Loss function에 L1 정규화를 추가한 기법인데, 중요하지 않은 B(계수)를 0으로 만들어 모델의 복잡성을 줄인다.
L2 정규화(Ridge) : 이상치(Outlier)의 Bi(계수)를 0에 가깝게 만든다.
-> Ridge Regression : Loss function에 L2 정규화를 추가한 기법인데, 중요하지 않은 B(계수)를 0에 가깝게 만들어 모델의 복잡성을 줄인다. Lasso와 다르게 완전한 0이 아니므로, 모델이 계속 복잡할 가능성도 있다.
Elastic Net Regression : Lasso, Ridge의 단점을 보완하기 위해 L1 정규화와 L2 정규화를 적절히 섞은 기법이다.

첫 번째로는 교차 검증(Cross Validation)이다. 훈련용 데이터와 별개의 테스트 데이터, 검증 데이터로 나누어 성능을 평가하는 방식이다. 일반적으로는 k-fold 교차 검증을 많이 사용한다.
k-fold 교차 검증은 쉽게 말해 데이터를 K등분하고 K번 훈련시키는 방법이다. 과정은 아래와 같다.

K를 설정하여 데이터 Set을 K개로 나눈다(K등분한다).
K개 중 한 개를 valid(검증)로 사용하고 나머지를 훈련용으로 사용한다.
K개 각각 모델의 평균 성능이 최종 모델의 성능이 된다.

feature importance를 고려하여 특성별로 A/B test 를 진행하며 feature selection 하시는 것을 추천

그러나 몇 가지 상황에서는 데이터를 정규화하는 것이 도움이 될 수 있습니다:

다른 유형의 모델과의 통합: 만약 그래디언트 부스팅을 다른 스케일에 민감한 모델과 함께 사

용한다면, 일관된 데이터 스케일링을 유지하는 것이 좋습니다.
이상치의 영향 최소화: 이상치가 있는 경우, 이상치의 영향을 줄이기 위해 데이터를 정규화할 수 있습니다.
학습 속도 향상: 때때로, 특성의 스케일을 정규화하면 모델의 학습 속도가 빨라질 수 있습니다.
결론적으로, 그래디언트