[A5] AutoML

2021312882 이준서

1. 설계한 AutoML pipeline 및 search space

(1) AutoML pipeline:

세 개의 데이터셋 모두 회귀 문제이므로, Pipeline의 step은 preprocessing과 regression으로 구성했습니다.

(2) Search space:

기본적인 전처리 없이 Autogloun을 이용해 모델별 성능을 알아보고, 성능이 높은 RandomForestRegressor, KNN Regressor, MLPRegressor을 사용했습니다. 이를 통해 기본적인 성능 Baseline을 설정할 수 있었습니다.

preprocessing: [None, StandardScaler, MinmaxScaler]

하이퍼파라미터는 { KNN의 n_neighbors: [3,5,7] , distance metric의 p: [1,2] / RandomForest의 n_estimators: [100, 300, 500], max_depth:[None, 10 , 30] / MLPRegressor의 'regressor__hidden_layer_sizes': [(50,), (100,), (100, 50)],'regressor__learning_rate_init': [0.001, 0.01, 0.1] }을 search 했습니다.

2. 각 데이터셋 별 최적의 모델 상세 정보 및 test set에서의 성능 평가 결과

Airfoil Data의 경우, 타겟의 분포가 정규분포와 하기에 RMSE를 사용해 평가했습니다.
Concrete Data의 경우 타겟의 분산이 크고 값의 범위가 넓기 때문에, MAE를 이용해 평가했습니다.
Abalone Data의 경우 타켓의 단위가 '년'이기에, 오차의 민감도를 줄이기 위해 MAE를 이용했습니다.

(1) Airfoil Data:

최적의 모델 및 하이퍼파라미터 조합: RandomForestRegressor (n_estimators=500, max_depth= 30)
학습 당시 최고 RMSE: 1.9659, Test set의 RMSE: 1.8115

(2) Concrete Data:

최적의 모델 및 하이퍼파라미터 조합: RandomForestRegressor(n_estimators=300, max_depth= None)
학습 당시 최고 MAE: 3.7032, Test set의 MAE: 3.7151

(3) Abalone Data:

최적의 모델 및 하이퍼파라미터 조합: MLPRegressor (hidden_layer_sizes=(100, 50), learning_rate=0.001)
전처리 방법: Standard scaling
학습 당시 최고 MAE: 1.5195, Test set의 MAE: 1.5365

3. 추가 성능 개선 및 AutoML 효율성 개선 방안 논의

모델과 하이퍼파라미터의 선택지를 줄이고자 AutoML을 먼저 사용하고 탐색을 진행했었습니다. 그러나 사용되는 모델의 다양성이 높지 않았고, 하이퍼파라미터의 기본값을 사용하기 때문에 큰 효율성을 얻지 못했습니다. CatBoost, WeightedEnsemble,NeuralNetTorch 등 다른 모델을 사용해본다면, 성능을 개선할 여지가 있을 것 같습니다.

도메인 지식을 반영하여 파생변수를 생성하는 등의 방법을 통해 성능을 개선할 수 있을 것 같습니다.

지금까지 배웠던 피처 엔지니어링을 이용해 성능 개선을 생각해볼 수 있을 것 같습니다. 가령 가우시안 분포를 따르지 않는 피처 분포들의 경우, log나 exp를 취하여 피처 전처리를 해줄 수 있을 것 같습니다. 또한 계산 성능이 더 높은 기기를 사용한다면, 간결한 코드로 많은 하이퍼파라미터와 조건들을 탐색함으로써 성능을 개선할 수 있을 것 같습니다.