[A5] AutoML

2021312882 이준서

1. 설계한 AutoML pipeline 및 search space

세 개의 데이터셋 모두 회귀 문제이기에, Pipeline의 step은 preprocessing과 regression으로 구성했습니다. 평가지표를 결정한 Autogloun을 이용하여 Baseline 성능을 잡아보았으며, 성능이 높은 모델들과 하이퍼파라미터를 중점적으로 search space를 탐색했습니다. Autogloun의 평가지표에서 대부분 부스팅 계열과 자체 앙상블 모델의 성능이 높게 나타났습니다. 이를 제외하고 성능이 높았던 모델인 RandomForestRegressor, KNN Regressor, MLPRegressor를 사용하였습니다.

- preprocessing은 상황에 맞게 [None, StandardScaler, MinmaxScaler]을 이용했습니다.

- 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.

Airfoil Data: RandomForestRegressor 모델을 사용하고, n_estimators=500, max_depth= 30 일 때 RMSE의 성능이 1.9659로 가장 좋았으며, Test set의 RMSE는 1.8115이었습니다.
Concrete Data: RandomForestRegressor 모델을 사용하고, n_estimators=300, max_depth= None 일 때 MAE의 성능이 3.7032로 가장 좋았으며, Test set의 MAE는 3.7151이었습니다.
Abalone Data: MLPRegressor 모델을 사용하고, StandardScaler()를 이용해 스케일링하고, hidden_layer_sizes=(100, 50),learning_rate=0.001일 때의 MAE 성능이 1.5195로 가장 좋았으며, Test set의 MAE는 1.5365이었습니다.

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.

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