**Assignment 5: Feature Engineering**

**2016314726 정영준**

**Jupyter notebook 환경에서 프로그래밍을 하였습니다. 사용한 데이터셋은 California housing 데이터셋으로 20640개의 row와 8개의 column으로 이루어져 있으며 각 feature는 모두 continuous value였습니다. Label의 경우 연속형 값으로 분류가 아닌 회귀를 적용하여야 하는 데이터셋입니다. 먼저 sklearn에서 dataset을 load하고 각 feature의 이름을 확인하였습니다. ‘MedInc, HouseAge, AveRooms, AveBedrms, Population, AveOccup, Latitude, Longitude’로 주택에 관한 정보가 feature로 지정되어 있습니다. Train set과 test set로 나누기 위해 numpy의 permutation을 활용하여 데이터의 80%를 train set으로 20%를 test set으로 나누었습니다. feature\_engineering이라는 함수를 정의하여 binning, polynomial transform, nonlinear transform을 한 함수를 사용하여 처리가 가능하도록 하였습니다. 사용한 머신러닝 모델은 선형모델 중 하나인 Ridge 모델을 사용하였습니다. Hyperparameter의 튜닝은 default값으로 통일하였습니다. 회귀의 metrics로 많이 사용되는 R2 score를 평가지표로 채택하였습니다. 먼저 아무 전처리를 하지 않은 데이터로 모델을 학습시키고 train set과 test set에 대해서 R2 score를 계산한 결과 train set에서 0.3516, test set에서 0.3572를 기록하였습니다. 모델의 비교를 위해 RandomForestRegressor 모델을 사용하여 얻은 test set의 0.7717 보다 매우 낮은 결과가 나왔습니다. 현재 데이터를 선형 모델로 근사하는 것이 쉽지 않다는 것을 알 수 있습니다. 선형 모델은 scaling의 영향을 크게 받기 때문에 Minmax scaler와 Standard scaler를 적용하여 다시 모델을 훈련시켰으나 비슷한 결과가 나왔습니다. 다음으로 Binning을 적용하였습니다. Test set의 데이터의 분포를 사용하지 않고 bins를 생성하고 이를 이용하여 train set과 test set을 구간 별로 categorical한 데이터로 변환하였습니다. 총 10개의 구간으로 나누었습니다. 그 후 R2 score를 측정한 결과 test set에서 0.6160으로 크게 모델의 성능이 향상되었습니다. 선형 모델에서 target값에 log를 적용시키는 것은 target의 분포를 변화시켜 성능을 향상시킬 수 있습니다. Log를 label에 적용한 결과 test set에서 0.6638의 R2 score를 얻을 수 있었습니다. 다음으로 degree를 6으로 설정한 polynomial 변환 후 Standard scaling, log target을 적용한 결과 test set에서 0.6249의 R2 score를 얻을 수 있었습니다. 다음으로 nonlinear transformation을 사용하여 log(x+1)을 x\_train에 적용한 결과 scaling, log target을 순차적으로 적용함에 따라 test set에서 R2 score가 0.4001, 0.5009, 0.5224로 증가하였습니다. 마지막으로 Feature selection을 사용하여 R2 score를 구하였습니다. SelectKBest에 f\_regression, mutual\_info\_regression을 적용한 결과 각각 0.3723 0.3227의 R2 score를 얻을 수 있었습니다. RFE의 경우 0.3723을 기록하였습니다. SelectFromModel을 사용하여 4개의 feature로 R2 score 0.3244를 얻을 수 있었습니다. Polynomial 변환과 Scaling이 이루어진 x값에 SelectFromModel을 적용한 결과 feature 3002개 중 1184개가 선택이 되었고 R2 score는 0.6221로 절반 이하의 feature를 사용하면서 큰 성능의 차이가 없는 feature selection이 이루어졌습니다. Binning으로 구간별 categorical feature로 변환하고 target값에 log를 적용한 Ridge 모델의 R2 score가 가장 높게 측정되었습니다.**

















