[4주차 과제2] KNN의 하이퍼파라미터 변경에 따른 성능결과 분석

자동차 가격 예측을 통한 분석

18011480 이진형 I 컴퓨터공학과

문제 분석

문제에서 주어진 데이터 셋(train.csv, test.csv)은 9개의 컬럼으로 구성되어 있다. 자동차 정보에 해당하는 데이터 셋으로 KNN 모델을 학습시킨 후, 자동차 가격을 예측하는 회귀 문제이다.

실험

데이터 셋 중 차량 모델명(model)이 실수화가 되어있지 않아 LabelEncoder를 통해 실수화 시켰고, StandardScaler를 통해 데이터 표준화를 진행했다.

이번 문제 역시 학습시킬 수 있는 데이터 셋의 크기가 작고, kaggle에 제출 후 정확도를 바로 확인할 수 있어 train_test_split을 진행하지 않아도 된다는 판단 하에 k와 weight값 두 가지만을 조정해 하이퍼파라미터 튜닝을 진행했고, 다음과 같은 결과를 얻었다.

weight	k	Score
uniform	4	1379.20488
uniform	3	1385.90664
uniform	5	1385.92063
uniform	6	1395.01420
uniform	7	1406.37429
uniform	8	1418.44103
uniform	2	1418.52646
uniform	9	1433.84815
uniform	1	1529.02857

weight	k	Score
distance	6	1332.92857
distance	5	1333.28670
distance	7	1334.99266
distance	4	1335.11301
distance	8	1338.14642
distance	9	1344.67217
distance	3	1350.72649
distance	2	1396.02807
distance	1	1529.02857

kaggle상 의 score에 따라 결과를 분석한 결과 weight = distance, k = 6 일 때 가장 높은 정확도로 모델이 만들어 진 것을 알 수 있다.

개선사항

추후 과제를 진행하면서 kaggle을 통해 바로 점수를 확인할 수 없을 때는 학습 데이터와 실험 데이터를 분리해서 정확도를 측정하는 과정이 필요한데, 이번 과제에서 진행했던 것 처럼 직접 튜닝하는 것 보다 scikit-learn 패키지의 GridSearchCV 모듈을 통해 최적화된 값을 도출해 낼 수 있다는 것을 알게 되어 추후 과제 진행시 사용하면 튜닝이 용이할 것 같다.