데이터마이닝 과제3 – 2022920024 박동찬

1. Linear Regression/Ridge Regression/Lasso Regression에서 hyper-parameter 탐색
   A. Linear Regression은 모델을 학습할 때 OLS(Ordinary Least Squares)을 사용하는데 기본적으로 정규화 hyper-parameter가 없다. 따라서 별도의 hyper-parameter tuning 과정을 거치지 않았다. 모델을 생성하고 train set에 대해 학습시킨 후, test set에 대해 예측하여 RMSE를 계산한다.
   B. Ridge Regression은 L2 정규화를 사용하여 가중치 학습에 규제를 준다. 이 경우에서 정규화 강도를 조절하는 alpha 값이 hyper-parameter로써 작용한다. Max_iter은 1000으로 설정하고 Alpha 값은 10^{-15}부터 10^{10}까지 10000개의 후보 값을 생성한다. GridSearchCV를 이용해서 각 alpha의 후보 값에 대해 모델을 학습시키고 5-fold cross validation(Default)을 수행한다. 그 중에서 가장 좋은 alpha 값을 선택하고 test set에 대한 예측과 RMSE를 계산한다.
   C. Lasso Regression은 L1 정규화를 사용하여 가중치 학습에 규제를 준다. 불필요한 feature의 계수를 0으로 만드는 효과를 가지고 있다. 이 경우에서도 정규화 강도를 조절하는 alpha 값이 hyper-parameter로써 작용한다. Alpha 값은 10^{-10}부터 10^{10}까지 10000개의 후보 값을 생성한다. GridSearchCV를 이용해서 각 alpha의 후보 값에 대해 모델을 학습시키고 5-fold cross validation(Default)을 수행한다. 그 중에서 가장 좋은 alpha 값을 선택하고 test set에 대한 예측과 RMSE를 계산한다.
2. 가장 좋은 모델의 coefficient, intercept, RMSE 값
   A. Linear Regression의 결과

```
----- Linear Regression -----
Intercept: -36.08420578532064
Coefficients: [ 4.33780990e-01  9.86011456e-03 -1.00451688e-01  5.99612449e-01
 -2.74446710e-06 -3.31740924e-03 -4.16479978e-01 -4.25956851e-01]
RMSE: 0.7248040193616437
```

   B. Ridge Regression의 결과

```
----- Ridge Regression -----
Best alpha: 78.8296628233039
Intercept: -35.684009811299255
Coefficients: [ 4.25706238e-01  1.00279439e-02 -8.48049508e-02  5.17478855e-01
 -2.09403547e-06 -3.30122294e-03 -4.13879613e-01 -4.22090824e-01]
RMSE: 0.7258014625763344
```

   C. Lasso Regression의 결과

```
----- Lasso Regression -----
Best alpha: 0.0022015789970454627
Intercept: -35.68186151345988
Coefficients: [ 4.25847367e-01  1.00182721e-02 -8.50611402e-02  5.20294265e-01
 -2.09210611e-06 -3.28600748e-03 -4.13812687e-01 -4.22036018e-01]
RMSE: 0.7257616627264676
```

3. Lasso에서 0이 되지 않은 weight에 해당되는 features
   A. 모든 feature(MedInc, HouseAge, AveRooms, AveBedrms, Population, AveOccup, Latitude, Longitude)의 계수가 0이 아니었다.
   B. MedInc는 중위 소득으로 주택 가격 예측에서 가장 중요한 변수로써 계수의 값이 약 0.4258로 나타났다. 소득이 높을수록 비싼 집을 구매하는 것과 같은 경향이 나타나고 있다.
   C. HouseAge(주택 연령)의 계수는 약 0.01002로 집값에 아주 미세한 영향을 미침을 알 수 있다. 양의 상관관계로 시간이 지남에 따라 주택의 가치가 변할 수 있다. 집 값과 크지 않은 관계를 가진다고 볼 수 있다.
   D. AveRooms(평균 방 수)의 계수는 약 -0.08506으로 영향이 크진 않지만 음의 상관관계를 가져 방의 수가 많아져도 집 값이 커지지 않음을 의미한다. 방의 수에 따라 집의 구조와 효율성에 영향을 미쳐 음의 상관관계가 나타날 수 있다.
   E. AveBedrms(평균 침실 수)의 계수는 약 0.52029로 가장 큰 양의 계수를 가지며 평균 침실 수에 따라 집 값이 크게 달라진다는 것을 알 수 있다. 평균 침실 수에 따라 주택의 크기와 거주 환경과도 큰 영향이 있다고 예측해볼 수 있다.
   F. Population(인구 수)의 계수는 약 -2.0921e-06으로 거의 0에 가까워 집 값과 크게 연관이 없다.
   G. AveOccup(평균 점유율)의 계수는 약 -0.003286으로 매우 작은 계수를 가지므로 집 값과 크게 연관이 없다.
   H. Latitude(위도)의 계수는 약 -0.4138으로 집 값에 중요한 영향을 미치며 음의 상관관계에 의해 위도가 작을수록 집 값이 증가하는 경향을 볼 수 있다.
   I. Longitude(경도)의 계수는 약 -0.4220으로 집 값에 중요한 영향을 미치며 음의 상관관계에 의해 경도가 작을수록 집 값이 증가하는 경향을 볼 수 있다. 즉, 위치에 따라 집 값이 달라진다는 것을 알 수 있다. 좋은 지역일수록 더 집 값이 비싸지는 경향을 보여주고 있다.
   J. Lasso Regression에서 선택된 best alpha 값이 매우 작기 때문에 모든 feature가 0이 아닌 계수를 가졌다. 하지만 집 값에 MedInc, AveBedrms가 높은 양의 관계를 가지고 Latitude, Longitude는 높은 음의 관계를 가져 이 features가 집 값과 큰 관계가 있음을 알 수 있다.

세 모델(Linear Regression/Ridge Regression/Lasso Regression) 모두 RMSE 값이 약 0.7258 정도로 거의 동일하게 나타난다. Regularization을 하더라도 Linear Regression 모델의 성능과 차이가 없는 것을 볼 수 있고 이는 California Housing dataset 자체가 비교적 선형적인 구조를 가지고 있음을 알 수 있다.