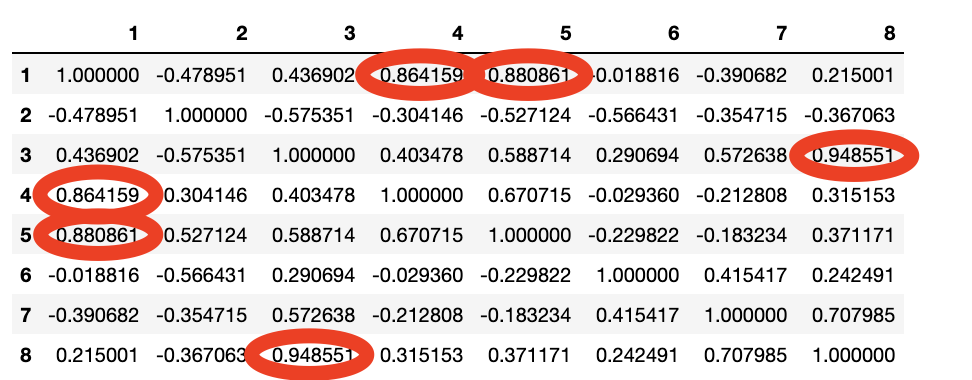
과목 : 기계학습 및 데이터마이닝

학과 : 소프트웨어학과

학번 : 201820742

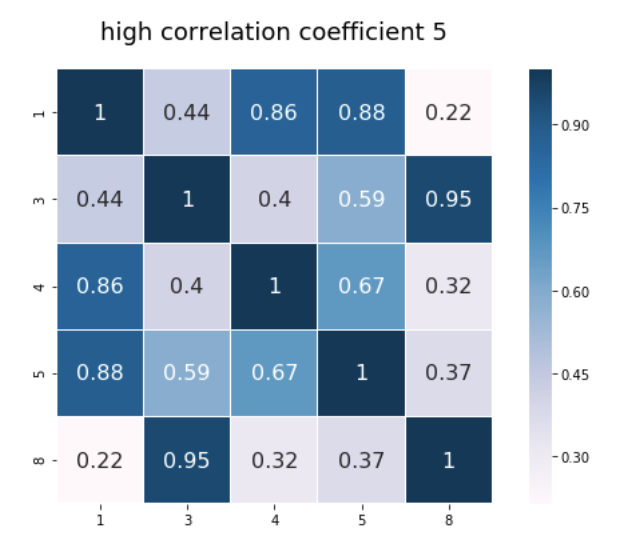
이름 : 고예준

1. **Coefficient interpretation (logistic regression)**
   1. 주어진 데이터의 모든 feature에 대해서 correlation coefficient array를 구하고 이 중 correlation coefficient 가 높은 5개의 feature를 이용해 feature간의 correlation을 **heatmap으로 그린다.** [5]



correlation coefficient array를 출력한 후, correlation coefficient가 높은 feature 5개를 뽑으면 1,3,4,5,8이라는 것을 알 수 있다.

이를 통해, heat map을 그려본다면 아래와 같은 결과를 얻을 수 있다.



* 1. 위에서 구한 correlation coefficient array를 바탕으로 multiple **logistic** regression 모델에서 사용할 feature를 정하고 90%의 데이터를 학습데이터로 사용하여 regression 모델 학습 후 10% 테스트 데이터에 대한 **test acc를 출력**한다. [5]



위에서 구한 correlation coefficient array를 바탕으로 사용할 feature인 1,3,4,5,8 이다. x\_train data와 x\_test data에서 해당 feature만을 골라내고, x data들을 정규화 시킨 후, multiple logistic regression 모델을 훈련시킨다. feature를 골라내고 정규화시킨 x\_train , 그리고 y\_train 값으로 훈련시켰다. 이는 분류 모델이므로 모델에 feature를 골라내고 정규화시킨 x\_test, 그리고 y\_test를 넣었을 때의 test acc를 구해주었다. 내 모델에서 test data에 대한 test acc는 0.94가 나왔다는 것을 알 수 있었다.

1. V**ariable selection (linear regression)**
   1. Linear regression 모델의 variable 선택을 통한 모델을 비교하기 위해서 수업시간에 배운 Adjusted 을 기반으로 forward selection을 수행한 후 **test 를 출력하고** x축은 선택한 variable의 개수, y축은 테스트 데이터의 Adjusted 과 학습데이터의 Adjusted 를 각각 다른 색으로 **그래프를 이용해 출력한다**.[10]
      * forward selection 을 이용하여 variable을 4개 까지 선택한다. (for 문 이용)
      * sklearn의 r2\_score 함수에 기반하여 Adjusted 를 만든다.

지도이(가) 표시된 사진

자동 생성된 설명

california data를 이용하여 x\_train, x\_test, y\_train,y\_test를 나누고, feature가 추가 됨에 따라 테스트 데이터의 Adjusted 과 학습데이터의 Adjusted 를 각각 다른 색으로 그래프를 그려주기 위하여 선택된 feature 수에 따른 train\_adjusted\_R^2와 test\_adjusted\_R^2를 list에 저장하도록 하였다. adjusted R^2 값을 구하기 위하여 adj\_r^2를 구해주기 위한 함수를 정의하였다.

forward selection을 구현하기 위해서 while문을 이용하여, feature가 4개까지 뽑히도록 하였고, while 문 안에서 for문으로 뽑힌 feature와 뽑히지 않은 feature 각각의 adj\_r2 score를 구하여 가장 score가 좋은 것을 하나씩 추가하여 뽑는 작업을 하였다.

이와 같은 과정으로 뽑힌 4개의 Feature는 ‘MedInc’,’HouseAge’,’Latitude’,’Longitude’였다. forward selection으로 변수 4개를 뽑은 후의 test data로 r^2 score를 구하면 약 0.5779이 나오고, adjusted r^2 score를 구하면 약 0.5771가 나오는 것을 확인할 수 있었다.

변수 개수에 따른 train adjusted r^2 값, test adjusted r^2 값을 구하면 위의 그래프 형태가 나온다는 것을 알 수 있다.

1. **Regularized linear model (linear regression)**
   1. Regularization 방법 중 하나인 lasso regression을 이용하여 5-fold cross validation을 통해 최적 하이퍼파라미터 탐색 후, 최종 모델의 test 를 출력한다. [10]
   * Scikit learn의 라이브러리를 기반하여 수행한다.

스크린샷이(가) 표시된 사진

자동 생성된 설명

내가 설정한 파라미터 alpha의 범위는 [0.001,0.01,0.1,1,10,100,1000]이었다. 5-fold로 cross-validation을 수행한 결과, 가장 높은 R^2값을 가진 곳은 alpha가 0.001인 경우였다. alpha가 0.001인 경우로 linear\_model.Lasso를 trian data로 훈련시키고 x\_test와 y\_test를 이용하여 r^2 value를 구한 결과, test R^2는 약 0.5904 이 나온다는 것을 알 수 있었다.

\* 주어진 모든 문제에 대해서 전체의 10% 데이터를 test set으로 두어 이 데이터에서의 성능을 test accuracy/로 출력한

\* lasso in sklearn : linear\_model.Lasso