PLAYGROUND SERIES SEASON 3 EPISODE 1

TABULAR REGRESSION WITH CALIFORNIA HOUSING DATASET

PREPARED BY: Dwaipayan Mukherjee (Emp. Id. 2211569)

Prepared as part of Kaggle projects given by Mr. Mrityunjoy Pandey.

The dataset for this competition (both train and test) was generated from a deep learning model trained on the California Housing Dataset. Feature distributions are close to, but not exactly the same, as the original. The dataset contains 9 feature columns, namely, Id, MedInc, HouseAge, AveRooms, AveBedrms, Population, AveOccup, Latitude and Longitude. The target column is the MedHouseVal. The data was available as csv files from the Kaggle site and came as a training dataset, ‘train.csv’, a testing dataset, ‘test.csv’, and a sample submission dataset, ‘sample\_submission.csv’.

Steps that I have followed:

1. I have first loaded the training data as a Pandas DataFrame and done some Exploratory Data Analysis to understand the nature of the dataset. The data set had no missing values in it. The population column was not in scale with the other data and so I have used a simple logarithmic transformation to the population column.
2. I have first tried to fit a Ridge Regression model with cross-validation to find the optimal parameter. With the best parameter value, the RMSE came out to be 0.77.
3. Next, I have used a Lasso Regression model with cross-validation to find the optimal parameter. With the best parameter value of 0.1, the RMSE came out to be 0.80, even worse than the ridge model.
4. Next, I used both the Ridge and the Lasso with their best parameters respectively along with a 10-fold cross-validation on the training dataset. The Average RMSE for the Ridge model came out as 0.89, while that for the Lasso model came out to be 0.93.
5. I used the Random Forest Regression model next along with a 10-fold cross-validation. The number of estimators was set to 1000. The Average RMSE came out to be 0.5899, quite an improvement on the previous algorithms.
6. Next, I used three boosting algorithms, namely, AdaBoost, XGBoost and LightGBM along with the 10-fold cross-validation on the training dataset.
7. The AdaBoost Regressor was initialized with number of estimators as 1000, and learning rate as 0.01. The Average RMSE came out as 0.9465. So, I increased the number of estimators to 10000, but the Average RMSE increased slightly to 0.9496.
8. The XGBoost Regressor was initialized with number of estimators as 10000, maximum depth as 9, learning rate as 0.01, colsample bytree as 0.66 and subsample as 0.9. The resultant Average RMSE after 10-fold cross-validation on the training dataset was 0.5584, the least till now. I increased the number of estimators to 20000 with other parameters same, and found the Average RMSE to be the same.
9. Lastly, I used the LightGBM Regressor with learning rate as 0.01, max depth as 9, number of leaves as 90, colsample bytree as 0.8, subsample as 0.9, subsample frequency as 5, min child samples as 36, reg lambda as 28, number of estimators as 20000 and metric as RMSE. The resultant Average RMSE after the 10-fold cross-validation on the dataset was 0.5587, just a little bit more than the XGBoost model.
10. So, I chose the XGBoost model as the final model for predictions and made predictions on the testing dataset.