Citation: Sklearn Documentation

1) **Statistical Analysis and Data Exploration**
    a) Number of data points (houses)?

        506 houses

    b) Number of features?

        13 features

    c) Minimum and maximum housing prices?

        Minimum housing price: 5k

        Maximum housing price: 50k

    d) Mean and median Boston housing prices?

        Mean: 22.53k

        Median: 21.2k

    e) Standard deviation?

        9.18 k

2) **Evaluating Model Performance**

    a) Which measure of model performance is best to use for predicting Boston housing data and analyzing the errors? Why do you think this measurement most appropriate? Why might the other measurements not be appropriate here?

        The mean squared error is the best in this situation. The mean squared error is basically a techinique that minimizes the average of the squared error (the difference of the truth and predicted values). It gives larger differences more weight than smaller differences. The mean squared error is a lot more capable in penalizing large errors than the mean absolute error technique. According to our statistics above, the minimum and maximum value are not outside the interval of -8.2 - 50.0, so there are no outliers. If there are outliers, the mean absolute error will be better because it is robust to outliers.

    b) Why is it important to split the data into training and testing data? What happens if you do not do this?

        It is important to split training and testing data because when you test on already seen data, and if your model is doing well, you will not be able to evaluate whether the model is actually doing well or whether it is overfitting. Overfitting will cause the classifier to just memorize the data that was trained

on and classify accurately only on foreseen data. It will perform badly in predicting new data because the classifier was not trained properly to learn the general pattern.

c) Which cross validation technique do you think is most appropriate and why?

A 10 fold cross validation is better than just randomly shuffling the sample and splitting the test and training data by 1 to 4 ratio. If you have a small amount of data, a k fold CV would have a lower variance than a hold out method. With a hold out method, you could be testing on a small test data, where there could be a lot of difference in performance for different samples of data. On the other hand, k fold validation reduces the performance variance by averaging the k different partitions.

d) What does grid search do and why might you want to use it?

According to sklearn documentation, grid search can set parameters that were not directly learned within the estimator by searching a parameter space of the best CV score. It can be useful for examples that have incomplete label data.

3) Analyzing Model Performance

a) Look at all learning curve graphs provided. What is the general trend of training and testing error as training size increases? (**EDITED)**

As training size increase, the test error decreases, and the training error gradually increases. For the test error, you can see a steep change from 0 to 40, then the error slowly decreases from 50 and so on. For the train error, the error slowly increases and stays flat from 200 and so on.

b) Look at the learning curves for the decision tree regressor with max depth 1 and 10 (first and last learning curve graphs). When the model is fully trained does it suffer from either high bias/underfitting or high variance/overfitting?

Max depth 1 is suffering underfitting because both the training and testing error are high.

As you increase the model complexity and is fully trained, the model at max depth 10 is suffering from overfitting. You see a higher variance between the training and testing error in max depth 10.

c) Look at the model complexity graph. How do the training and test error relate to increasing model complexity? Based on this relationship, which model (max depth) best generalizes the dataset and why?

The training error gets better as the model complexity increases, but the testing error gradually worsens. Max depth of 6 best generalizes the data set because there is a higher error rate for max depth 0 - 5 where it suffers underfitting, and there is a higher variance between the training and testing error for max depth greater than 6 where it suffers overfitting.

4) Model Prediction

a) Model makes predicted housing price with detailed model parameters. Compare prediction to earlier statistics

Best model parameter: {'max_depth': 6}

House: [11.95, 0.0, 18.1, 0, 0.659, 5.609, 90.0, 1.385, 24, 680.0, 20.2, 332.09, 12.13]

Prediction: [20.76598639]

This is a reasonable prediction because the prediction is close to the median housing price value of 21.2 and well within the standard deviation.