Question 1

Of the features available for each data point, choose three that you feel are significant and give a brief description for each of what they measure.

Answer:

CRIM: This variable measures the per capita crime rate of the town. I would expect houses in areas with a larger per capita crime rate to be less valuable, all other things being equal.

PTRATIO: This variable measures the pupil to teacher ratio of the town. Less pupils per teacher can indicate better schools, and many people prefer to live in neighborhoods with better schools for their kids. I would therefore expect houses in areas with a larger pupil to teacher ratio to be less valuable, all other things being equal.

RAD: This variable is an "index of accessibility to radial highways". I would expect houses in areas with easier access to radial highways to be more valuable, all other things being equal.

Question 2

Using your client's feature set `CLIENT_FEATURES`, which values correspond with the features you've chosen above?

Answer: *CRIM* is the first feature in the data set - for the client, this value is 11.95. *PTRATIO* is the eleventh feature, with a value of 20.2 for the client. *RAD* is the ninth feature, with a value of 24 for the client.

Question 3

Why do we split the data into training and testing subsets for our model?

Answer: Splitting the data means we can perform tests on data that wasn't used to fit the model. This is important to verify that the model is adequately predicting values for new data. If we test the model with the same data used to train it, we won't know if the model is overfitting - that is, fitting very precisely for the training set, thus increasing the variance of the model and consequently the error rate for new data.

Question 4

Which performance metric below did you find was most appropriate for predicting housing prices and analyzing the total error. Why?

- *Accuracy*
- *Precision*
- *Recall*
- *F1 Score*
- *Mean Squared Error (MSE)*
- *Mean Absolute Error (MAE)*

Answer: *MSE*. Since we are dealing with a regression problem, metrics for classification (accuracy, precision, recall and F1 score) should not be used. Of the two metrics used for regression problems, I chose the Mean Squared Error because it penalizes the model more for larger errors. I should be more worried by 1 error of value 10 than by 10 errors of value 1; with MSE, the first case will give me an error of 100 while the second one will give me an error of 10 (with MAE, both cases would give me an error of 10).

Question 5

What is the grid search algorithm and when is it applicable?

Answer: Grid search algorithm fits and tests a model using different parameters (for example, the degree of a polynomial function, the maximum depth of a decision tree, or the number of nearest neighbors). This allows us to choose the best set of parameters from all the sets defined in the grid. Grid search algorithm is applicable whenever we must estimate "hyperparameters" - that is, parameters that must be set *before* the model is fit to the training data.

Question 6

What is cross-validation, and how is it performed on a model? Why would cross-validation be helpful when using grid search?

Answer: Cross-validation means dividing the data into k different folds and using each fold in turn as the testing set. We train the model k times, each time withholding a different fold ans using it to test our model. As a result, we have k different test error rates, which can be averaged to calculate a final test error rate for the model.

When using grid search, cross-validation can be used to estimate the test error rate for the model when using each set of parameters defined in the grid. The set of parameters with the smallest test error rate should be used to fit the model.

Question 7

Choose one of the learning curve graphs that are created above. What is the max depth for the chosen model? As the size of the training set increases, what happens to the training error? What happens to the testing error?

Answer: For max depth **1**, the training error increases and the testing error decreases as the number of data points in the training set increases. Both training and test errors seem to converge somewhere between 40 and 60.

Question 8

Look at the learning curve graphs for the model with a max depth of 1 and a max depth of 10. When the model is using the full training set, does it suffer from high bias or high variance when the max depth is 1? What about when the max depth is 10?

Answer: With a max depth of **1**, the training and testing errors are similar and somewhat high, which indicates a high bias model (the model cannot approximate the "true function" that governs the data). With max depth **10**, the training error is very low but the testing error is high, which indicates a high variance model (the model captures all variation in the training set almost perfectly, but it's unable to give good predictions for new data because it fitted the noise in the training set along the way).

Question 9

From the model complexity graph above, describe the training and testing errors as the max depth increases. Based on your interpretation of the graph, which max depth results in a model that best generalizes the dataset? Why?

Answer: max depth of 4. The testing error decreases until we reach max depth of 4, and then stabilizes. Therefore, a model with a max depth of 4 is the simplest model we can use that will give us the smallest verified testing error, and we should use it.

Question 10

Using grid search on the entire dataset, what is the optimal `max_depth` parameter for your model? How does this result compare to your intial intuition?

Answer: The optimal *max_depth* parameter is 4, which agrees with my initial intuition.

Question 11

With your parameter-tuned model, what is the best selling price for your client's home? How does this selling price compare to the basic statistics you calculated on the dataset?

Answer: The predicted value for the client's home is 21.630 thousand dollars. It is well within the range of values in the dataset, between the median and the mean calculated above.

Question 12 (Final Question):

In a few sentences, discuss whether you would use this model or not to predict the selling price of future clients' homes in the Greater Boston area.

Answer: With 10 folds, I calculated the cross-validation error of the model to be approximately 36. Since this is the mean *squared* error, I took the square root to get to an average error of approximately 6 thousand dollars. This average error is 2/3 of the standard deviation of approximately 9 thousand dollars, which seems acceptable, although I think it may be improved by a different model.